Implementation of PCM (Process compact models) for the study and improvement of variability in advanced FD-SOI CMOS technologies

Yvan Denis

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Nanoélectronique et nanotechnologies (NENT)**

Arrêté ministériel : 7 août 2006

Présentée par

Yvan Denis

Thèse dirigée par **Gérard Ghibaudo** et
Co-encadrée par **Frédéric Monsieur**

préparée au sein du **Laboratoire IMEP-LAHC**
dans **l'École Doctorale EEATS**

Implémentation de PCM (Process Compact Models) pour l'étude et l'amélioration de la variabilité des technologies CMOS FD-SOI avancées

Thèse soutenue publiquement le 16/06/16,
devant le hury composé de :

Pr. Francis Calmon
Professeur des Universités INSA, Président

Dr. Damien Deleruyelle
Maître de conférences à Aix-Marseille Université, rapporteur

Pr. Pascal Masson
Professeur à l'Ecole Polytechnique Universitaire de Nice Sophia-Antipolis, rapporteur

Dr. Thierry Poiroux
Docteur au CEA-LETI Grenoble, examinateur

Pr. Gérard Ghibaudo
Directeur de recherche à l'IMEP-LaHC, directeur de thèse

Dr. Frederic Monsieur
Docteur à STMicroelectronics Crolles, encadrant de thèse

Communauté
UNIVERSITÉ Grenoble Alpes

THESIS

In order to obtain the degree of

DOCTOR OF GRENOBLE ALPES UNIVERSITY

Specialty : **Nanoelectronic and nanotechnologies (NENT)**

ministerial decree: 7th august 2006

Presented by :

Yvan Denis

Thesis supervised by **Gérard Ghibaudo** by
Co-conducted by **Frédéric Monsieur**

Prepared within **Laboratoire IMEP-LAHC**
In the **École Doctorale EEATS**

Implementation of PCM (Process compact models) for the study and improvement of variability in advanced FD-SOI CMOS technologies

Thesis publically defended the 16th, june 2016.
In front of the following dissertation committee:

Pr. Francis Calmon
Professor at Universités INSA, President

Dr. Damien Deleruyelle
Distinguished lecturer at Polytech' Marseille, reporter

Pr. Pascal Masson
Professor at l'Ecole Polytechnique Universitaire de Nice Sophia-Antipolis, reporter

Dr. Thierry Poiroux
PhD researcher at CEA-LETI Grenoble, reviewer

Pr. Gérard Ghibaudo
Professor at l'IMEP-LaHC, member

Dr. Frederic Monsieur
PhD researcher at STMicroelectronics Crolles, member

Table of Contents

Table of acronyms

AIC	Akaike Information Criterion
BOx	Burried Oxide
BPV	Backward Propagation of Variance
df	degree of freedom
DIBL	Drain Induced Barrier lowering
DOE	Design Of Experiment
DSA	Dynamic Surface Anneal
F	Fisher value
f dose	Implant energy and dose multiplicative factor
Fcrit	Critical Fisher value
FDC	Fault Detection and Classification
FD-SOI	Fully Depleted Silicon On Insulator
FFNN	Fead Forward Neural Network
FPE	Final Prediction Criterion
FPV	Forward Propagation of Variance
GAA	Gate All-Around
k-fold CV	k-fold Cross-Validation
KNN	Kernel Nearest Neighbor
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LDR	Lightly doped region
LER	Line Edge Roughness
LFN	Low frequency Noise
LOOCV	Leave One Out Cross-Validation
MC	Monte Carlo
MGG	metal gate granularity
MOS	Metal Oxide Semiconductor
MOSFET	Metal On Semiconductor Field Effect Transistor
MSE	Mean Square Error
nMOS	n type Metal Oxide Semiconductor
OLS	Ordianary Least Square
PCM	Process Compact Model
PD-SOI	Partially Depleted Silicon On Insulor
pMOS	p type Metal Oxide Semiconductor
POR	Process Of Reference
PSG	polysilicon granularity
PT	Parametric Tests
Qhk	Defect density at the high-K layer interface
RDD	Random Discrete Dopant
Rext	External resistance added to the contact
RMS	Root Mean Square Error
SCE	Short Channel Effects
SE	Standart Error

SOI	Silicon On Insulator
SPC	Statistical Process Control
SRAM	Static Random Access Memory
SSE	Sum of Square Error
SSR	Sum of Square Regression
STI	Shallow Trench Insulation
SVM	Suport Vector Machine
TCAD	Technology Computer Assisted Design
TED	Transcient enhanced diffusion
Tepi	Epitaxial layer thickness
Til	Insulating layer thickness
Tsi	SOI layer thickness
Tspike	Spike anneal temperature
UTB	Ultra Thin Body
UTBB	Ultra Thin Body and Box
UV	Ultra-Violet
Wsp	Spacer width

Table of notations

β	Gain factor
β_i	ith polynomial coefficients
$\epsilon(0^+)$	maximum potential barrier height at the virtual source
ϵ_{ox}	Silicon oxide permitivity
ϵ_s	Silicon permitivity
ϕ_b	Difference between the intrinsic silicon Fermi level and quasi Fermi level of doped silicon
ϕ_m	Metal work function
ϕ_{MS}	Metal-Silicon built-in Voltage
ϕ_s	Silicon work function
ϕ_{th}	Surface potential at threshold
σ	Vg dependant access resistance in linear regime
σ_e	electrical parameter covariance matrix
σ_m	model parameter covariance matrix
θ_1	First order high field mobility reduction factor
θ_2	Second order high field mobility reduction factor
φ_s	surface potential at the Si/SiO2 interface
φ_{sb}	surface potential at the Si/Box oxide interface
φ_{sf}	surface potential at the Si/Gate oxide interface
$\hat{Y}$	Observations
μ	Carrier mobility
μ_0	universal mobility
μ_{eff}	Carrier effective mobility
C_{sub}	Capacitance induced by the depletion layer formed at the Box substrate interface
C_{BOx}	BOx capacitance
C_{ox}	Gate oxide capacitance
D_{it}	Defect concentration at the gate interface
E_{eff}	Effective electric field experienced by the carrier in the channel
E_c	conduction band energy
$E_{f,metal}$	metal fermi energy
$E_{f,silicon}$	silicon fermi energy
E_g	Silicon band gap
E_i	intrinsic energy
E_{lat}	lateral field
E_{ox}	Electric field across the gate oxide
E_s	electrical field at the silicon/gate oxide interface
E_{sb}	electrical field at the silicon/BOx oxide interface
E_{sf}	electrical field at the silicon/gate oxide interface
E_v	valence band energy
E_{vacuum}	vacuum energy
G_m	Transconductance
I_{Din}	Linear drain current
I_{Dsat}	Saturation drain current

J	Jacobian matrix
k_B	Boltzmann constant
L	Physical gate length
Lc	critical length at which µ_short is half the long channel µ mobility
L_{eff}	effective channel length
m^*	Carrier effective mass
m_t	transverse electron mass
M_v	product of the lowest valley degeneracy and the reciprocal of the fraction of the carrier population in the lowest energy level
Na	Acceptor concentration
Q	Charge of the electron
q_1	25th percentile
q_3	75th percentile
Q_{acc}	Accumulation charge density
Q_d	Depletion charge
Q_i	inversion carrier charge
Q_{ib}	inversion carrier charge at the back interface
Q_{if}	inversion carrier charge at the front interface
R_0	Constante access resistance in linear regime
R_{on}	Total transistor resistance in saturation regime
R_S	Source access resistance
R_{SD}	Source plus drain access resistance
R_{tot}	Total transistor resistance in linear regime
T	temperature
T_{BOx}	BOx thickness
$T_{dep\text{-}BOx}$	Depletion layer thickness at the silicon/BOx interface
$T_{dep\text{-}Ox}$	Depletion layer thickness at the silicon/gate interface
T_{ox}	Gate oxide layer thickness
T_{si}	SOI layer thickness
v^*	carrier velocity limitation due to vinj and vsat
V_B	Substate bias
V_C	quasi Fermi potential along the channel
V_D	Drain bias
V_{Dsat}	Drain bias at which the drain current saturates
V_{FB}	Flatband voltage
V_{FBB}	Flat band voltage at the back interface
V_{FBF}	Flat band voltage at the front interface
V_G	Gate bias
V_{inj}	injection velocity
V_{ox}	Potential drop in the oxide of the MSO Structure
V_S	Source bias
v_{sat}	saturation velocity
V_t	Threshold voltage
V_{tLDR}	Lightly doped region threshold voltage
V_{tlin}	Threshold voltage in linear regime

V_{tsat}	Threshold voltage in saturation regime
W	Transistro width
X	Variables or predicators
Y	Model response

Chapter 1 :
Introduction

Moore's law predicted that it will take two years to step from a technological node to the next one. Even though this pace has been kept until recently, the trend starts to slow down as declared by Intel CEO in 2015 [1]. Thus, pursuing MOSFET scaling trend becomes harder as the time passes. This loss of speed is due to many factors. First, the length associated with the technological node is supposed to reflect the average half-pitch of a memory cell, which is the size of a pattern in an array of transistor used to build memory cells. In fact, what it has been observed between the 90 and the 30 nm nodes is that each time a node is crossed, the area of the chip is scaled by ½, but the gate pitch is actually scaled by 0.7 and the physical gate length by only 0.9. In the meantime, source-drain junction optimization succeeded to reduce sufficiently the overlap distance so that the effective channel length did not change at all between the 90 nm down to the 30 nm technology node as shown in Figure 1-1 [2].

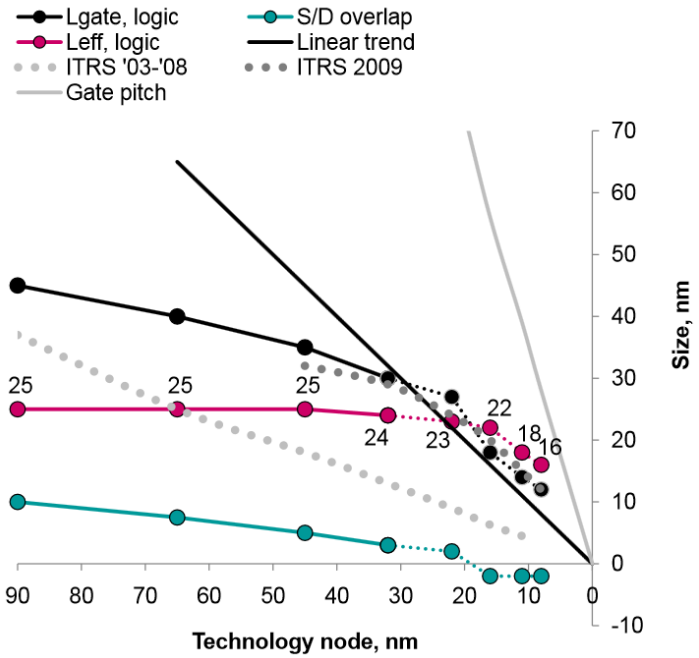


Figure 1-1: Transistor Size Evolution: ITRS 2009

So by reducing the gate pitch and the gate length with a proper source drain junction engineering, it has been possible to keep the same effective length (L_{eff}) down to the 30 nm node. However, in order to go forward, this condition could not hold and L_{eff} had to be reduced, inducing new challenges such as increased short channel effect due to the loss of control of channel's electrostatic. These challenges are part of the reason why the development time is longer.

In addition variability is increasing and becomes a greater challenge. Indeed, the number of variability sources and their impact increase as shown by R. Sitte comparing 1.5 µm and 0.1 µm bulk MOSFETs [3][4], by Fu-Liang Yang comparing 45, 32, 22 and 16 nm technology node [5] and by Samar K. Saha using ITRS roadmap [6][21]. However performance uniformity of elementary devices like transistors is a priority for microelectronic manufacturers. Indeed, any dispersion in this performance will be propagated on the next circuit level (e.g. SRAM). From one circuit level to the other the impact of performance dispersion is often increased. As a consequence, a small dispersion of performance at the lowest architecture level can jeopardize the highest circuit level functionality. Large dispersions lead to large yield loss and an increase in the circuit production cost. This is why huge effort is done to bring MOSFET performance dispersion down to a reasonable range.

Considering variability, its sources are differentiated depending on their autocorrelation length and their statistical nature [14][15]. In particular we distinguish global from local variability (considering the autocorrelation length) [16][17][25] and statistical from systematic variability (considering the statistical nature) [18]:

- Global variability encompasses every sources of variability that has a larger autocorrelation length than die dimension. Hence, it comprises within-wafer, wafer-to-wafer, lot-to-lot and across-factory variability (e.g. deposited insulating layer thickness across the wafer, anneal temperature …).
- Local variability only encompasses within-die variability sources (e.g. Random Discrete Dopant (RDD), Line Edge Roughness (LER), transistor orientation with respect to crystalline orientation, well proximity effect)…
- Systematic variability arises from sources that can be predicted (e.g. transistor orientation with respect to crystalline orientation, well proximity effect …).
- Statistical variability, in contrast with systematic variability, is characterized by its stochastic nature. It can only be comprehended using statistical modeling (e.g. anneal temperature, Random Discrete Dopant, Line Edge Roughness…).

Typical source of local random variability are RDD, LER, polysilicon and metal gate granularity (PSG/MGG), Trapped Interface Charges (TIC) and interface roughness. Local random variability can hardly be controlled by process adjustment because of its stochastic nature. Indeed corresponding sources are stochastic phenomena that impact each device independently one from another at the atomistic scale. Since these phenomena are random, it is neither possible to accurately predict their impact at the device level nor to counteract them by process optimization. Only stochastic variability model can be used to predict the dispersion of the transistor's performance. Even though a lot of work has been done to limit its impact by circuit design optimization [19][20], this source is intrinsic to the technology used. This is why it is considered as the bottom line in terms of variability. Thus, adopting new architectures or using advanced process techniques like extreme UV are the only ways to reduce local random variability. Consequently, a large amount of work has been dedicated to systematically investigate local random variability sources for every technology node [21]-[29].

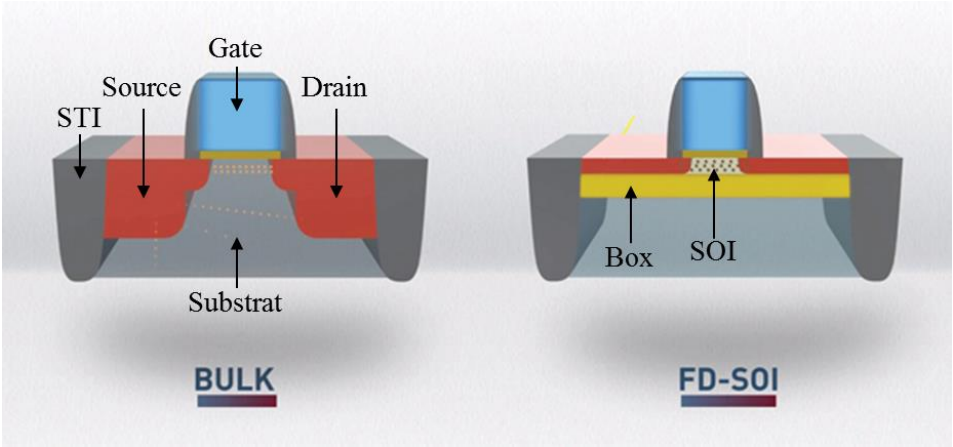


Figure 1-2: Comparison between Bulk and FD-SOI architecture.

In order to meet both local random variability and gate length reduction requirement, new architectures and techniques have been introduced like Fully Depleted Silicon on Insulator (FD-SOI) [10], FinFet [11], double gate [12] or Gate All-Around (GAA) MOSFETs [13], extreme UV etching,

SiGe and III-V components. As an example, we show the case of FD-SOI MOSFET in Figure 1-2 comparing the FD-SOI and bulk architecture.

This architecture meets the gate length reduction requirement by improving significantly the electrostatic control of the channel thanks to the buried oxide layer (BOx) beneath the channel. To another extend, it enables controlling the threshold voltage by tuning the back interface voltage, enabling low power applications. FD-SOI technology addresses local random variability issue using intrinsic channel. In addition, it has been shown to be less sensitive to Line Edge Roughness (LER) [21]. So this technology enables reducing significantly the impact of local random variability. Due to these advantages, it has been chosen by STMicroelectronics for the 28 nm technological node and beyond.

Global variability also limits the yield. Indeed, large process dispersion at wafer or lot scale can compromise the functionality of a large number of die. Thus process and device performance should be monitored at the die scale. In order to fulfill this task, performance indicators have been determined. These indicators are continuously controlled with in-line measurements called Parametric Tests (PT). A reasonable dispersion of performance is then obtained if every one of these indicators lie within predetermined boundaries called Statistical Process Control (SPC). Performance indicators are chosen as critical quantities that will limit the next circuit level performance. For example cell current is a performance indicator for SRAM circuits. This cell current is directly related to the saturation current of the transfer (access) transistor and pull-down (driver) transistor [7][8]. This is why SPC have been created for the saturation drain current (I_{Dsat}). Another example is the frequency of ring oscillator at operating voltage. This frequency is linked to the switching speed of CMOS inverters that are themselves related to the peak current obtained during inverter switching, commonly defined as the average between I_H and I_L, where $I_H=I_D(V_G=V_{DD}, V_D=0.5V_{DD})$ and $I_L = I_D(V_G=0.5V_{DD}, V_D=V_{DD})$ [9]. V_G, V_D and V_{DD} are the gate, drain and operating voltage respectively. I_D is the drain current. Consequently SPC includes I_H and I_L as well. Similarly to performance control, process control is carried continuously using Fault Detection and Classification (FDC). This technique continuously monitors equipment parameters against preconfigured limits using statistical analysis techniques to provide proactive and rapid feedback on equipment health.

However new architectures and techniques are more complex solutions. It requires more process steps and thus more photo-lithography masks. Extreme UV lithography requires specific tools that are more expensive. These options are very different from the tradition way used to build transistors. Increased process complexity also induces larger global variability. For example, the SOI thickness variability, absent in bulk architecture, is a new contribution to the global variability. All these facts tend to increase the time and investment required to develop and optimize the next generation of device. In order to limit the development cost and ensure the profitability and competitiveness of these new devices it is mandatory to rely on more efficient approach for the device development and optimization at industrial stage. This thesis aims at offering a determinist and robust approach able to meet these expectations.

Indeed its goal is to demonstrate how it is possible to model the relationships between process parameters (accessible to tool engineers) and the transistor performances. Such a model is called Process Compact Model (PCM). A sufficiently robust and predictive PCM can be used for optimizing the performance and global variability of the transistor thanks to an appropriate optimization algorithm. This approach is different from traditional methods that heavily rely on expert knowledge and successive trials in order to improve the device since it brings a deterministic and robust mathematical frame to the problem.

The task is not trivial and faces many constraints. First, there are hundreds of process steps required only for the front-end-of-line and at least as much process parameters can be distinguished from this. This implies dealing with a large amount of data. Thus robust and adapted statistical tools are required to manage this issue. Second, the physical relation between process and electrical parameters are complex. Many models describing the MOSFET electrical parameters exist in the literature, but since we deal with a large amount of data and intend to use it for optimization, only simple compact analytical model can be used.

The PCM investigated in this thesis copes with these constraints. It is composed of two stages. Starting from process parameters, the first stage is formed of multiple polynomial formulas that relate process with the model parameters of a typical threshold based compact model. The second stage is the compact model. Using model parameters as inputs, it yields electrical parameters as output. An input/output scheme of this two-stage PCM is presented in Figure 1-3.

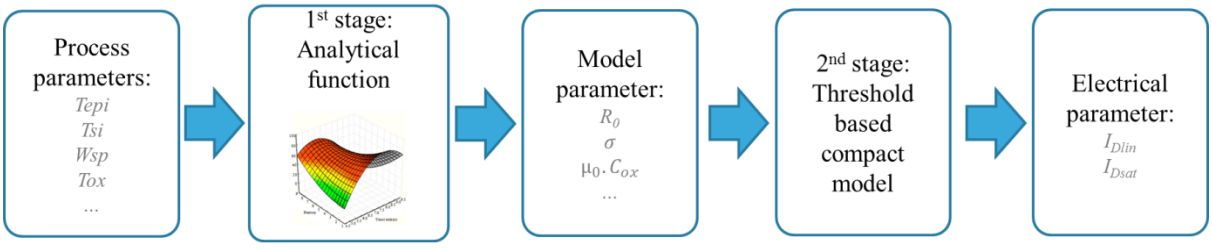


Figure 1-3: Scheme of the two-stage PCM

This manuscript starts by introducing the compact model used. Analytical formulation of linear and saturation drain currents are drawn from the physics of semiconductors. This stage aims at splitting a complicated global parameter that is drain current into simpler and physically meaningful sub parameters such as access resistance, threshold voltage, carrier mobility and so on. Theoretical derivation of the model is done in chapter 2 where the physics of the transistor is developed with an emphasis on the specificity of FD-SOI technology.

Using a compact model for drain current implies to rely on an extraction procedure in order to get model parameters and calibrate the model. An extraction procedure is proposed in chapter 3, based on a nonlinear optimization algorithm. The method robustness is tested against data sample size and range as well as the effect of noise. Then, in the same chapter, the extraction procedure is tested on a TCAD simulated Design Of Experiment (DOE). This DOE exhibits process parameters variations in order to investigate their effects on extracted model parameters. This is a first approach to examine the link that we miss yet to model the process and electrical parameters relations. The physical relevance of the model parameters sensitivity to process variation is demonstrated to ensure that model parameters are physically meaningful and that extraction procedure is robust.

Extractions are then performed on silicon (using 28 nm FD-SOI and 14 nm FD-SOI technologies) in chapter 4. Lots have been processed with various kinds on process variations. These lots have been measured and from these measurements, model parameters have been extracted. It is then shown how results can be interpreted to give insights into the device characteristics. While being a simple approach, this method can already produce valuable results and indicate how to optimize the device efficiently.

In order to complete the PCM construction, a map of the process and model parameters relationships are required. This topic is investigated in chapter 5. Model parameters are much more elementary than drain current. This advantage enables us to build simple model such as empirical polynomial model. However, process parameters are numerous and all of them have not necessarily a significant impact

on model parameters. Moreover, the impact of noise in measurements and local stochastic variability in devices induces an increased uncertainty in model parameters. This is why statistical methods are introduced in order to efficiently build polynomial model dealing with ill-posed problem and noisy observables. These methods are tested against synthetic data and applied on TCAD extractions to build PCMs. The impact of noise and local stochastic variability is discussed and solutions to deal with those issues are investigated.

Finally, based on this PCM, a methodology to optimize both electrical performances and variability is suggested. It has been applied using TCAD simulation to indicate how to reduce drain current global variability efficiently.

Chapter 2 :
Transistor’s drain current compact modeling

This chapter is devoted to detail the compact model used for the PCM. It derives the drain current equations, starting with the derivation of inversion carrier concentration and threshold voltage of Metal Oxide Semiconductor (MOS) capacitance in §2.1. Carrier mobility is investigated in §2.2. Paragraphs 2.3 and 2.4 introduce the linear and saturation drain current approach respectively. Access resistance effect is introduced in §2.5. The drain current equations derived here will be reused for the model extraction procedure detailed in Chapter 3.

2.1 The MOS capacitance and its electrostatics

In this section we derive the MOS capacitance equations of the inversion carrier charge Q_i and the threshold voltage (V_t) in the case of the Bulk structure. Then we discuss the case of Fully Depleted Silicon On Insulator (FD-SOI) with doped and undoped channel and show to which extent the same compact equations for Q_i can be used. Only the threshold voltage dependence should be adapted and we will show how and why. Notice that in further derivation, interface and oxide charges are neglected. However considering it would add only small corrections and the derivation would still hold.

2.1.1 Inversion in bulk MOS transistors

The MOS capacitance is first treated in order to derive the inversion charge density in the long channel bulk transistor. The corresponding band diagram is shown in Figure 2-1 where elements are isolated:

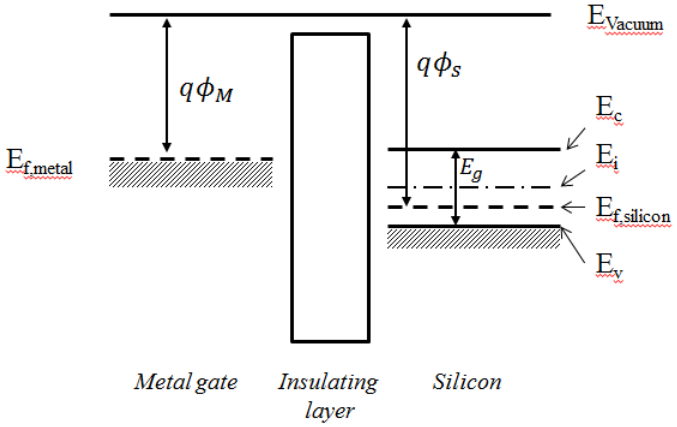


Figure 2-1: Energy band diagram of a classical MOS structure with a P doped silicon layer. Shaded area represents the electron populated energy levels.

Left part of the MOS structure in this figure is the metal gate. It is separated from the p-doped silicon bulk (right part) by an insulating layer (gate oxide). In Figure 2-1, $E_{f,metal}$, $E_{f,silicon}$, E_i, E_c, E_v and E_{Vacuum} are respectively the metal and silicon Fermi energy, the intrinsic, conduction band, valence band and vacuum energies for isolated parts. E_g is the semiconductor gap energy (E_c-E_v). ϕ_s and ϕ_M are the silicon and metal work functions. The gate and bulk biases are V_G and V_B. Every voltage are referenced to a hypothetical unbiased neutral body where there is no band bending. This reference potential is $E_{f,silicon}/q$ and, in this case, is equal to V_B. Later we will see that in the case of transistor, this potential reference depends on the position in the channel and is no longer equal to V_B. This diagram shows that, when taken separately, Fermi levels of metal and semiconductor are not matched.

Thus when building a MOS structure, charges (holes for p-doped silicon) will be repealed at the Si/SiO_2 interface, creating a potential drop across the insulator and a space charge region in the silicon

channel in order to reach the equilibrium state where Fermi levels of metal and silicon are aligned (see Figure 2-2).

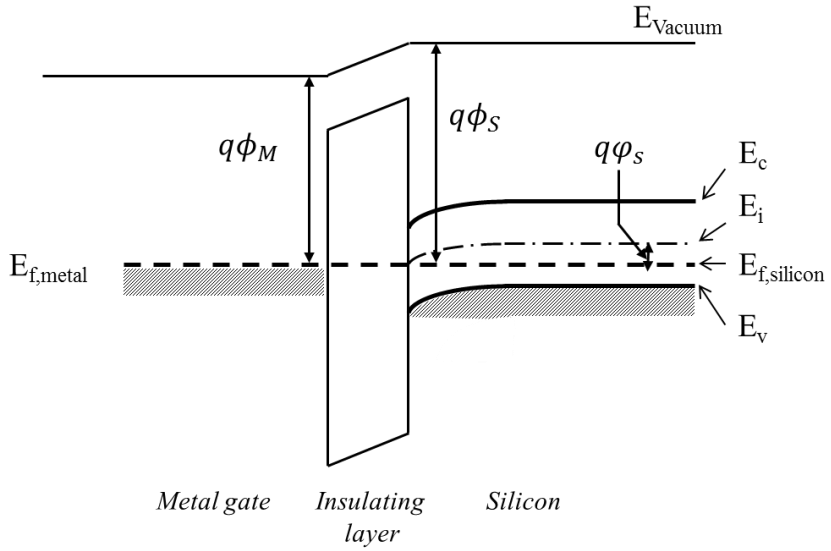


Figure 2-2: Energy band diagram of a classical MOS structure at equilibrium for grounded electrodes (metal and silicon)

Thus without applying any bias between the gate and the back electrode ($V_G = V_B = 0$), there already is a potential drop across the MOS structure that is equal to work function difference between metal and silicon ($V_{FB} = \phi_M - \phi_s$). Space charge region in silicon contains no free carrier but ionized dopants. The total charge density in this case is called the depletion charge density.

If now, for the case of p-doped silicon, a positive bias V_G between gate and bulk is applied, then $E_{f,metal}$ decreases and the bands bend even more until a certain point where minority carrier concentration equals majority carrier concentration at the Si/SiO_2 interface. This point characterizes the beginning of the inversion regime where free carriers appear at the Si/SiO_2 interface (see Figure 2-3 left). In that case the total charge density is the sum of inversion and depletion charge ($Q = Q_d + Q_i$).

On the contrary if a negative bias is applied to the gate, V_G tends to compensate the built-in voltage across the oxide capacitance ϕ_{MS} and the silicon bands bend less. If $V_G = V_{FB}$ then the MOS structure reaches the flat band condition (see Figure 2-3 right), no potential drop occurs across the structure and the channel is electrostatically neutral.

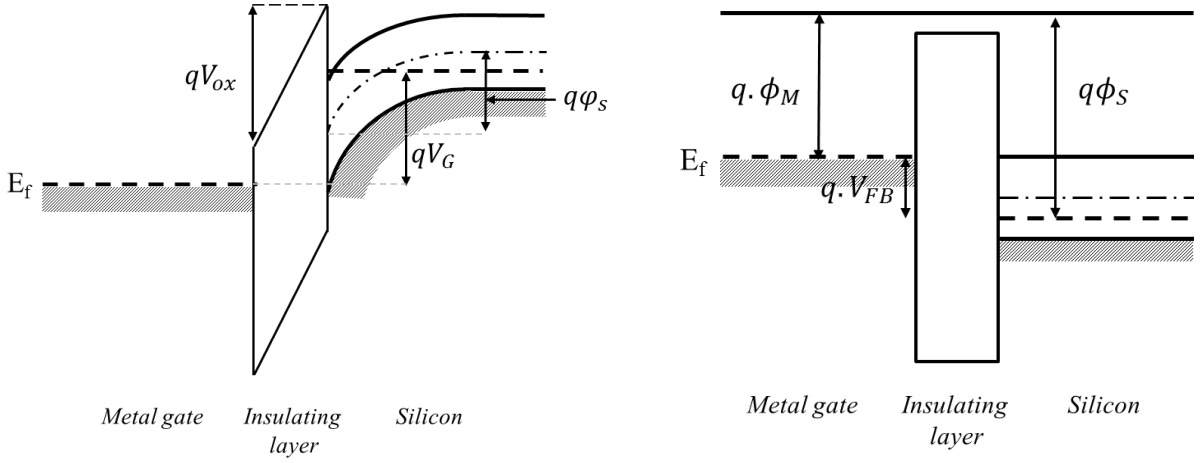


Figure 2-3: MOS band diagram in inversion (left) and flat band (right) regime

Driving further V_G toward negative values will accumulate holes at the Si/SiO_2 interface. This is the accumulation regime. This qualitative approach enables a global understanding of the relationship between V_G and carrier concentration in the channel. The quantitative and rigorous derivation of the carrier density depending on gate voltage and doping concentration in the channel has been done by R. H. Kingston and S. F. Neustadter [30]. However, here, we will use an alternative approach sufficiently accurate for our purpose. First the potential drop in the structure is the sum of the potential drop in the oxide V_{ox} plus the potential drop in the silicon [31]:

$$V_G - V_{FB} = V_{ox} + \varphi_s \tag{1}$$

where V_{FB} is the flat band voltage, φ_s is the surface potential at the Si/SiO_2 interface referenced to a hypothetical unbiased neutral body where there is no band bending (as depicted in Figure 2-2 Figure 2-3). The electric field in the oxide E_{ox} is constant because there is no \charge in the oxide:

$$E_{ox} = \frac{V_{ox}}{t_{ox}} \tag{2}$$

where t_{ox} is the insulating layer thickness. At the Si/SiO_2 interface, in the silicon but before reaching any charge, the displacement field is constant and the electrical field E_s, at the silicon interface is:

$$E_s = \frac{\epsilon_{ox} E_{ox}}{\epsilon_s} \tag{3}$$

Then we can write E_s as a function of V_G:

$$\epsilon_s E_s = \epsilon_{ox} \frac{V_{ox}}{t_{ox}} = C_{ox}(V_G - V_{FB} - \varphi_s) \tag{4}$$

where $C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$ is the gate oxide capacitance. In absence of any interfacial charges, using Gauss theorem over an area going from Si/SiO_2 interface up to $x = \infty$ in the silicon bulk where the reference is taken, we get the following relation:

$$-\epsilon_s E_s = Q_d + Q_i \tag{5}$$

Thus the inversion carrier density reads:

$$Q_i = -C_{ox}(V_G - V_{FB} - \varphi_s) - Q_d \tag{6}$$

The threshold is reached when the minority carrier concentration equals the majority carrier one. This is achieved when the band bending φ_s reaches $2\phi_b$ [32] where ϕ_b is the difference between the intrinsic silicon Fermi level and quasi Fermi level of doped silicon. From this we can deduce V_t from (6) where $Q_i = 0$:

$$V_t = V_{FB} + 2|\phi_b| + \frac{1}{C_{ox}}\sqrt{4\epsilon_s q N_a |\phi_b|} \tag{7}$$

Then Q_i above threshold is deduced from equation (6) where $\varphi_s = 2\phi_b$ and becomes:

$$Q_i \approx -C_{ox}(V_G - V_{FB} - 2|\phi_b|) + \sqrt{4\epsilon_s q N_a |\phi_b|} \tag{8}$$

It can be expressed in terms of V_t as:

$$Q_i \approx -C_{ox}(V_G - V_t) \tag{9}$$

2.1.2 Inversion in FD-SOI MOS transistors

FD-SOI denomination is employed for transistors built on SOI substrate and that has a fully depleted channel at operating condition. This is the kind of structure that is used by STMicroelectronics for their 28 and 14 nm technological nodes. In this structure the simple MOS capacitance cannot be used to calculate the charge density in the channel since the Buried Oxide (BOx) adds another capacitance contribution that shall be taken into account when applying gauss law. The following paragraph explains the SOI structure and derives the necessary conditions to have a FD-SOI structure. Then we will adapt previous derivation of inversion charge and threshold voltage to FD-SOI structure. The results will depend on the characteristics of the device. First, for the derivation of bulk MOS capacitance we consider a doped channel but state-of-the-art industrial FD-SOI devices have intrinsic channel. It implies that the surface potential is no longer clamped to $2\phi_b$ in strong inversion regime and that Q_d can be neglected. Second, FD-SOI devices built by STMicroelectronics are made of ultra-thin substrates and BOx. We will see that ultra-thin SOI induces front to back super coupling effect that does not allow inversion at one interface along with accumulation at the other. Finally, with ultra-thin BOx, we will see that the ground plane regime also affects the threshold voltage, depending on its doping concentration.

2.1.2.1 Structure presentation

The capacitance structure we consider is a cross section of the FD-SOI MOSFET in the middle of the channel. This structure is shown in Figure 2-4:

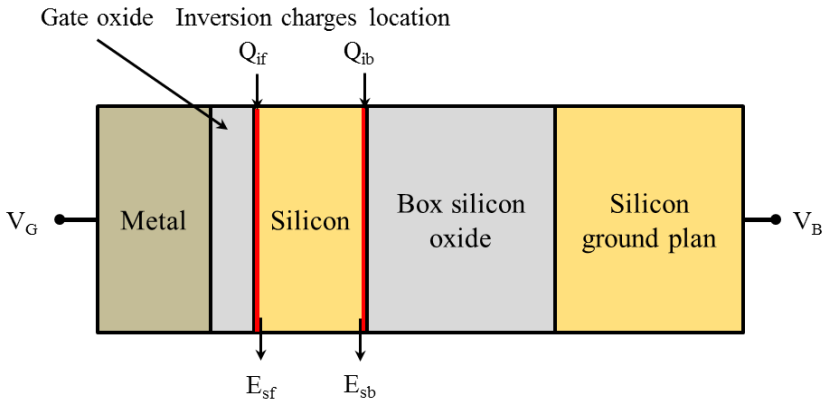


Figure 2-4: MOS SOI capacitance structure considered for inversion carrier concentration derivation. E_{sf} and E_{sb} are front and back Si/SiO_2 interfaces electric field respectively.

Basically we see that SOI capacitance is no more than one MOS and one semiconductor-oxide-semiconductor capacitance (assuming that gate and BOx oxide thicknesses can be different) assembled head to tail sharing the same FD substrate. Thus in the following we will talk about front and back interface to designate the gate oxide/SOI and the SOI/BOx interface respectively. Considering a thick SOI layer for the channel, corresponding band diagram is shown in Figure 2-5.

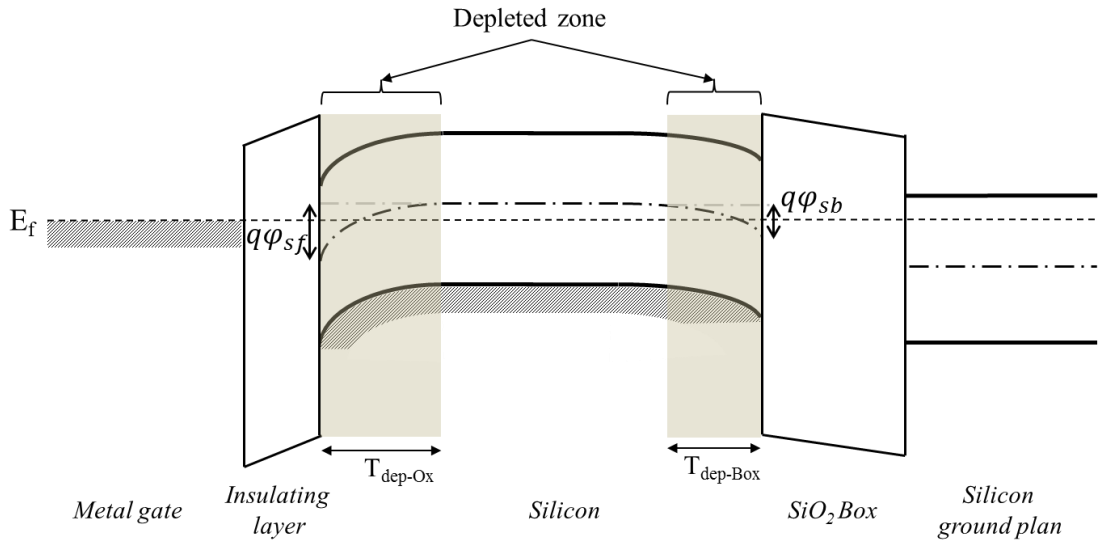


Figure 2-5: Band diagram of a partially depleted SOI MOSFET at equilibrium. φ_{sf} and φ_{sb} are the front and back interface potential respectively.

This is the case of Partially Depleted Silicon On Insulator (PD-SOI) MOSFET. In Figure 2-5 $t_{dep\ Ox}$ and $t_{dep\ BOx}$ are the maximum depletion thickness at the front and back gate. The difference between PD-SOI and FD-SOI lies in the silicon channel thickness. Indeed, in order to have a fully depleted silicon channel the silicon channel thickness (T_{si}) should be lower than $t_{dep\text{-}Ox}+t_{dep\text{-}BOx}$. Thereafter, we derive the critical channel thickness that differentiates PD-SOI from FD-SOI structure. $t_{dep\ Ox}$ and $t_{dep\ BOx}$ quantities are derived from the expression of the depletion charge [33]:

$$Q_d = -\sqrt{2\epsilon_s q N_a \varphi_s} \tag{10}$$

where N_a is the acceptor concentration. Let's focus on the front gate depletion thickness ($t_{dep\ Ox}$). From equation (2) to (6) we have the following relationship between V_{ox} and t_{dep} at onset of strong inversion (where Q_i=0):

$$V_{ox} = -\frac{Q_d}{C_{ox}} = \frac{\sqrt{qN_a 2\epsilon_s \varphi_s}}{Cox} = \frac{qN_a t_{dep\text{-}Ox}}{C_{ox}} \tag{11}$$

$$\varphi_s = \frac{qN_a t_{dep\text{-}Ox}^2}{2\varepsilon_{si}} \tag{12}$$

and between V_G and t_{dep}:

$$V_G = V_{FBF} + \varphi_s + V_{ox} \tag{13}$$

where $V_{FBF} = \phi_M - \phi_s$ is the flat band voltage of the MOS capacitance. On the other hand, the flat band voltage of BOx capacitance will be labeled V_{FBB}. Then replacing V_{ox} and φ_s in (13) by their expression in (10) and (11), t_{dep} is deduced as a function of V_G and the doping concentration:

$$t_{dep\text{-}Ox} = \frac{\epsilon_s}{C_{ox}}\left(\sqrt{\frac{2(V_G - V_{FBF})C_{ox}^2}{qN_a\epsilon_{si}} + 1} - 1\right) \tag{14}$$

$t_{dep\text{-}BOx}$ formulation is identical to (14) except that C_{ox}, V_G and V_{FBF} should be replaced by C_{BOx}, V_B and V_{FBB}. If $t_{si} < t_{dep\,Ox} + t_{dep\,Box}$, the unbiased region in the middle of the channel disappear and the device becomes fully depleted. Critical silicon thickness to have a fully depleted device is calculated from the following formula:

$$t_{si\text{-}crit} = t_{dep\text{-}Ox} + t_{dep\text{-}Box}$$

$$t_{si\text{-}crit} = \frac{\epsilon_{si}}{C_{ox}}\left(\sqrt{\frac{2(V_G - V_{FBF})C_{ox}^2}{qN_a\epsilon_s}} + 1 - 1\right) + \frac{\epsilon_{si}}{C_{BOx}}\left(\sqrt{\frac{2(V_B - V_{FBB})C_{BOx}^2}{qN_a\epsilon_s}} + 1 - 1\right) \quad (15)$$

$t_{si\text{-}crit}$ ranges from 3.1 µm down to 51 nm for $10^{16} < N_a < 10^{19}$ cm^3, $V_G = V_B = 1$V, $T_{ox} = 1.3$ nm, T_{BOx}=25 nm, $V_{FBF} = -0.52$ V and $V_{FBB} = 0.84$ V at room temperature. The ground plan doping has been set to 10^{19} cm^{-3}. Below this thickness, the back and front interfaces become coupled and the band diagram becomes more complex. Figure 2-6 shows the band diagram at $V_G = 1.5V$ and $V_B = [-6, 0, 6]V$.

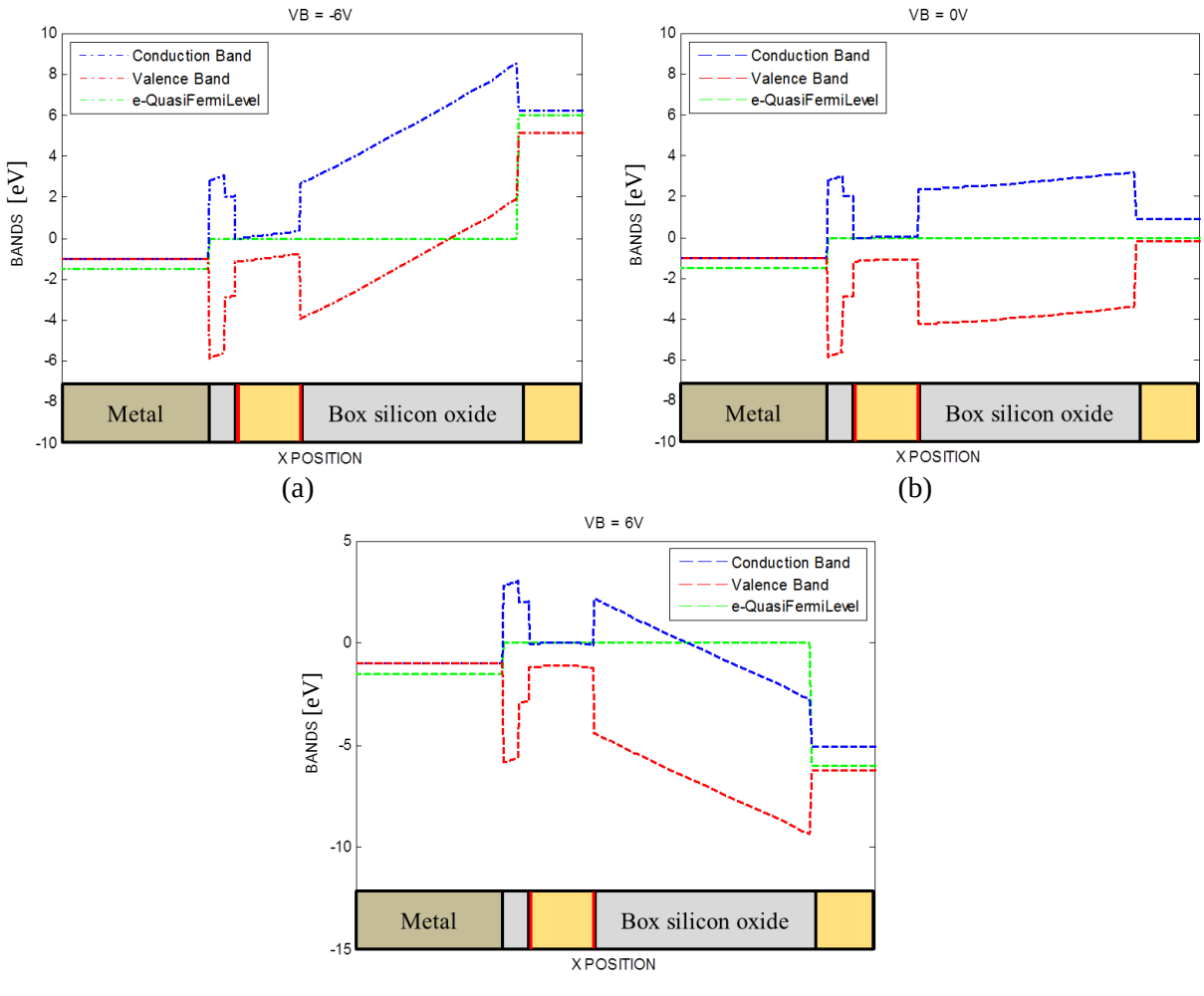


Figure 2-6: MOS SOI band diagram for $V_G = 1.5$V and $V_B = -6$V (a), $V_B = 0$V (b) and $V_B = 6$V (c). Band diagram has been generated using UTOXPP Poisson-Schrödinger solver [34].

We see from this plot that the back interface goes from depletion to inversion depending on V_B. Figure 2-7 shows the minority carrier concentration depending on the position in the SOI for many V_B

ranging from -10 up to 10V. In this case the SOI is lightly p-doped (5.10^{15} atoms/cm^3) and the BOx is 25 nm thick. V_G is still set to 1.5V.

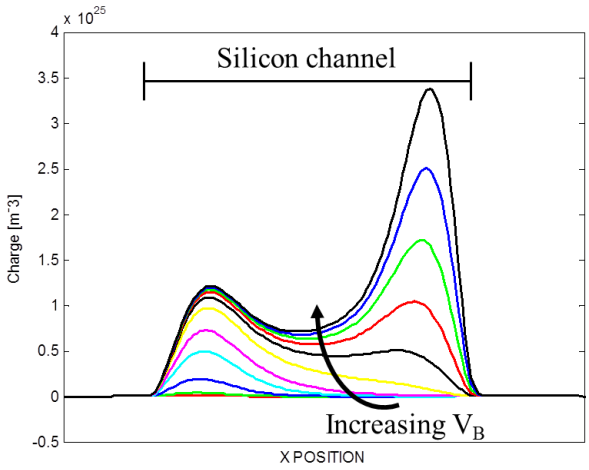


Figure 2-7: Minority carrier concentration in the SOI for V_B going from -10 up to 10V at $V_G = 1.5V$. Charge concentration has been calculated using UTOXPP Poisson-Schrödinger solver [34].

This figure shows that the inversion layer is not concentrated at the front interface as it was the case for bulk devices. Now we have to consider the possibility of having a back interface inversion layer induced by V_B. In any case we see that the electrostatics of one interface is influenced by the other one through back to front interface coupling.

2.1.2.2 Inversion charge density and threshold voltage derivation

In the following paragraph we will investigate the expression of Q_i and V_t for the case of back interface accumulation, depletion and inversion. Following derivations assume the delta-depletion approximation, that is, any inverted charge is at the Si surface in a Dirac delta function as depicted in Figure 2-4. In order to adapt bulk inversion charge density and threshold voltage derivation to FD-SOI structure, we consider a doped channel FD-SOI structure with thick SOI (but thin enough to deal with FD-SOI structure, that is SOI thickness < $t_{dep\ ox}$ as mentioned in (14)) and BOx. We will see in the next paragraphs the effect of thin BOx and channel and intrinsic channel. This investigation has been conducted by H-K Lim and J. G. Fossum [35]. Following the same approach as for bulk MOS capacitance we can derive the potential drop across both front and back oxide (as in (1)):

$$V_G - V_{FBF} = V_{ox} + \varphi_{sf} \tag{16}$$

$$V_B - V_{FBB} = V_{box} + \varphi_{sb} \tag{17}$$

where V_{FBF} and V_{FBB} are the flat band voltages of front and back gates respectively. φ_{sf} and φ_{sb} are the surface potentials at the front and back Si/SiO_2 interfaces respectively, referenced to a hypothetical unbiased neutral body. Solving the Poisson equation across the silicon film yields:

$$\frac{V_{si}}{t_{si}} = E_{sf} - \frac{qt_{si}N_A}{2\varepsilon_{si}} \tag{18}$$

where $V_{si} = \varphi_{sf} - \varphi_{sb}$ is the potential drop across the SOI film, E_{sf} is the field at the front interface of the SOI, t_{si} is the silicon film thickness, ε_{si} is silicon dielectric permittivity, N_A is the doping concentration in the channel. The field in the front oxide is:

$$E_{ox} = \frac{V_G - V_{FBF} - \varphi_{sf}}{t_{ox}} \tag{19}$$

Then, applying Gauss theorem across the front oxide, taking the inversion charge into account, yields:

$$\varepsilon_{ox} E_{ox} - \varepsilon_{si} E_{sf} = -Q_{if} \tag{20}$$

where Q_{if} is the inversion layer at the front side. Equation (18) to (20) can be adapted for the back interface. To sum up, the four following equations are available (following Lundström [36]):

$$E_{sf} = \frac{\varepsilon_{ox}}{\varepsilon_{si}} \frac{V_G - V_{FBF} - \varphi_{sf}}{t_{ox}} + \frac{Q_{if}}{\varepsilon_{si}} \tag{21}$$

$$E_{sb} = \frac{\varepsilon_{ox}}{\varepsilon_{si}} \frac{-V_B + V_{FBB} + \varphi_{sb}}{t_{box}} - \frac{Q_{ib}}{\varepsilon_{si}} \tag{22}$$

$$E_{sf} = \frac{\varphi_{sf} - \varphi_{sb}}{t_{si}} + \frac{q t_{si} N_A}{2\varepsilon_{si}} \tag{23}$$

$$E_{sb} = \frac{\varphi_{sf} - \varphi_{sb}}{t_{si}} - \frac{q t_{si} N_A}{2\varepsilon_{si}} \tag{24}$$

where Q_{ib} is the inversion layer at the back side. Combining (21) with (23) and (22) with (24) yields the general equations that rule the electrostatic of the SOI capacitance with doped SOI:

$$V_G = V_{FBF} + \varphi_{sf} - \frac{Q_{if} + \frac{Q_d}{2}}{C_{ox}} + \frac{C_{si}}{C_{ox}} (\varphi_{sf} - \varphi_{sb}) \tag{25}$$

$$V_B = V_{FBB} + \varphi_{sb} - \frac{Q_{ib} + \frac{Q_d}{2}}{C_{box}} + \frac{C_{si}}{C_{box}} (\varphi_{sb} - \varphi_{sf}) \tag{26}$$

where $C_{si} = \frac{\varepsilon_{si}}{t_{si}}$ and $Q_d = -qN_A t_{si}$ is the depletion-region areal charge density. Combining these two equations yields the back and front coupling equation. Then the total inversion carrier concentration is simply the sum of Q_{if} and Q_{ib}:

$$Q_i = E_{sf}\varepsilon_{si} - C_{ox}\left(V_G - V_{FBF} - \varphi_{sf}\right) - E_{sb}\varepsilon_{si} - C_{box}(V_B - V_{FBB} - \varphi_{sb}) \tag{27}$$

Replacing E_{sf} and E_{sb} by their expression (23)(24) yields:

$$Q_i = -C_{ox}\left(V_G - V_{FBF} - \varphi_{sf}\right) - C_{box}(V_B - V_{FBB} - \varphi_{sb}) - Q_d \tag{28}$$

Let's now discuss the expression of the threshold voltage. In order to derive it we need a clear definition of this threshold voltage. The subtlety we are facing here is that there are two gates. Thus considering one electrode with a fixed bias, the threshold voltage is the other electrode voltage required to switch the transistor from off to on state (or from on to off). Notice that this is not the definition used by Lim and Fossum [35]. Instead they defined the threshold voltage as V_G for which $\varphi_{sf} = 2\phi_b$ no matter the back interface regime. This definition fails to match ours when the back interface is inverted. However our definition reflects the threshold voltage that is extracted by most of the extraction procedure, which is not the case of Lim and Fossum definition.

A consequence of this definition is that at threshold, $Q_i = 0$. Thus considering V_B fixed, the threshold voltage can be deduced from V_G expression (25), depending on φ_{sb} and φ_{sf}, where $Q_i = 0$.

$$V_t = V_{FBF} + \varphi_{sf} - \frac{Q_d}{2.C_{ox}} + \frac{C_{si}}{C_{ox}}(\varphi_{sf} - \varphi_{sb}) \tag{29}$$

From this definition, Q_i expression can be simplified using V_t.

$$Q_i = -C_{ox}(V_G - V_t) \tag{30}$$

So we see that Q_i expression is similar to bulk one (9) except for the definition of V_t and is valid whatever the value of φ_{sb} and φ_{sf}, hence whatever V_B value. Next, V_t expression is developed, clarifying φ_{sf} and φ_{sb}, depending on V_B and whether the back interface is accumulated, depleted or inverted.

- Threshold voltage when the back interface is accumulated

In the case of an accumulated back interface, accumulated charges screen the back bias and φ_{sb} is virtually pinned at 0. Threshold at front interface is then reached when $\varphi_{sf} = 2\phi_b$ as in bulk case. Then Q_{if} is small and V_t is deduced from (25):

$$V_t = V_t^A = V_{FBF} + \left(1 + \frac{C_{si}}{C_{ox}}\right)2\phi_b - \frac{Q_d}{2C_{ox}} \tag{31}$$

Vt does not depend on the back potential in this case.

- Threshold voltage when the back interface is inverted

If the back interface is assumed inverted, then φ_{sb} equals $2\phi_b$ and a conducting back channel exists. The current flows if $V_G = 0$. Thus V_t is the required gate bias to suppress the back channel inversion. If when $V_G = V_t$, the front interface is depleted, φ_{sf} is deduced from (26):

$$\varphi_{sf} = 2\phi_b\left(1 + \frac{C_{box}}{C_{si}}\right) - \frac{C_{box}}{C_{si}}(V_B - V_{FBB}) - \frac{Q_d}{2C_{si}} \tag{32}$$

Then V_t is deduced from (25):

$$V_t = V_{FBF} + 2\phi_b - \frac{C_{box}(C_{si}+C_{ox})}{C_{ox}C_{si}}(V_B - V_{FBB} - 2\phi_b) - Q_b \cdot \frac{C_{ox}+2C_{si}}{2C_{si}C_{ox}} \quad (33)$$

If, when V_G equals V_t, the front interface is accumulated, then $\varphi_{sf} = 0$ and V_t is clamped to:

$$V_G = V_{FBF} - \frac{C_{si}}{C_{ox}} 2\phi_b \quad (34)$$

- Threshold voltage when the back interface is depleted

In the case of a depleted back interface, φ_{sb} depends on V_B and ranges from 0 up to $2\phi_b$ corresponding to the limit case of back interface accumulation and inversion respectively as we have seen previously. Isolating φ_{sb} from (26) and introducing it in (25), letting $Q_{if} = 0$ and $Q_{ib} = 0$, we get:

$$V_t = V_{FBF} + 2\phi_b - \frac{C_{si}C_{box}}{C_{ox}(C_{si}+C_{box})}(V_B - V_{FBB} - 2\phi_b) - \frac{Q_d}{2}\frac{2C_{si}+C_{box}}{C_{ox}(C_{si}+C_{box})} \quad (35)$$

- Threshold voltage summary depending on back bias

To conclude there are four cases to consider as reported by F. Andrieu [37] (see Figure 2-8):

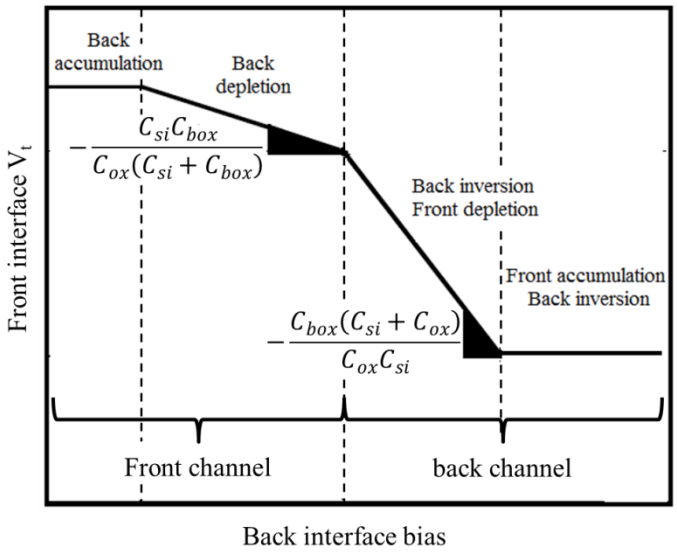


Figure 2-8: Theoretical V_t-V_B curve with the different channel regimes.

When the back interface is accumulated, V_t does not depend on V_B and is expressed as (31). When the back interface is depleted, V_t-V_B slope is $-\frac{C_{si}C_{box}}{C_{ox}(C_{si}+C_{box})}$ (see equation (35)). When the back interface is inverted, there are two options: i) either the front interface is depleted, then V_t is expressed as (33) and the V_t-V_B slopes shift to $-\frac{C_{box}(C_{si}+C_{ox})}{C_{ox}C_{si}}$ or ii) the front interface accumulated and then V_t does not depend on V_B (as expressed in (34)).

2.1.2.3 Effect of intrinsic instead of doped channel

In order to reduce random dopant fluctuations and enable a better scalability of FD-SOI technology [38], intrinsic channel has been preferred and is the actual option used by STMicroelectronics. The main difference is that the depletion charge Q_b is now almost zero and the depletion approximation that implies $Q_d \gg Q_i$ is not valid anymore. V. P. Trivedi et al. [39] have refined H-K Lim and J. G. Fossum [35] approach (developed in §2.1.2.2) and made it suitable for undoped channel. In particular, they showed that for subthreshold condition $\frac{|Q_i|}{C_{ox}}$ is smaller than $|\varphi_{sf}|$ and the influence of subthreshold or weak inversion charge on φ_{sf} can be neglected irrespective of the channel doping condition. Thus, simplification used thanks to depletion approximation still holds for intrinsic channels. The master equations (25) and (26) are identical except that Q_i and Q_d becomes negligible in weak inversion.

Another consequence of the lack of impurity in the channel is that the quasi Fermi level in silicon is close to the intrinsic Fermi level, dropping ϕ_b to 0. Thus at threshold, $\varphi_{sf} = \varphi_{th}$ is substantially greater than $-2\phi_b$ and a better definition of φ_{th} should be found. Lee and Young [40] and V. P. Trivedi et al. [39] have adapted Lim and Fossum [35] approach by defining a critical surface electron concentration n_T needed for the channel to be conductive. Then φ_{th} yields:

$$\varphi_{\mathrm{th}} = \varphi_0 - \phi_b \tag{36}$$

Where $\varphi_0 = -\phi_b$ if $N_A > n_T$ and $\varphi_0 = \frac{\mathrm{kT}}{\mathrm{q}} \ln\left(\frac{\mathrm{n_T}}{\mathrm{n_I}}\right)$ otherwise. While being simple and efficient, this approach requires to arbitrarily set $\mathrm{n_T}$ to a specific value. This approach is similar to the constant current threshold voltage definition and has been reported by Q. Chen et al. [41]. Another approach similar to maximum of transconductance criterion has been suggested by J. Lacord et al. [42] and yields:

$$\varphi_{\mathrm{th}} = \frac{\mathrm{kT}}{\mathrm{q}} \ln\left(\frac{\mathrm{kT}}{\mathrm{q}^2 \mathrm{n_i} \mathrm{T_{si}}} \mathrm{C_{ox}}\right) \tag{37}$$

if the inversion layer thickness is equal to the silicon film thickness. Here are two examples of threshold surface potential definition but many others have been published.

Beside the fact that $2\phi_b$ should be replaced by φ_{th} and that Q_d can be neglected, Trivedi et al. [39] showed that the general Vt expression is similar to the case of doped channel for depleted back interface.

2.1.2.4 BOx and channel thickness limiting effects: the case of Ultra-Thin Body and BOx (UTBB) structure.

Previous derivations hold if the back to front coupling is weak enough to enable inversion at one interface and accumulation at the other. However S. Eminente [43] showed that for ultra-thin t_{si}, due to the strong back to front electrostatic coupling, the required electric field in the BOx to induce accumulated charges at the back interface while having the front interface inverted would not be supported by the BOx oxide. Thus the back accumulation regime is not worth investigating. Figure 2-9a shows that even for thick BOx (that sustains higher voltage), considering front channel at threshold, back interface accumulation is reached only when V_B<-40V and back interface inversion

(with front interface accumulation) when V_B>60V. These voltages are very far from the operating voltage, making it unusable.

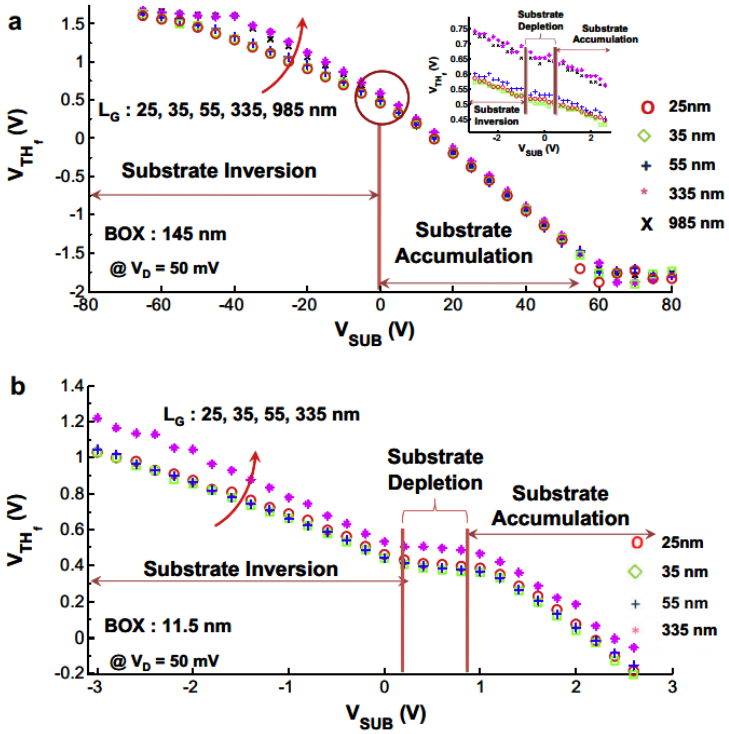


Figure 2-9: Threshold voltage extracted for different gate lengths as a function of the substrate bias voltage and for two BOX thicknesses: (a) for UTB and (b) for UTBB . V_B is varying from -3 V to +3 V for UTBB and from -80 V to +80 V for UTB.[44]

2.1.2.5 Effect of substrate depletion on threshold voltage

Another difference in the case of Ultra-Thin Body and BOx (UTBB) is the influence of substrate regime. In the case of ultra-thin BOx, it appears that the BOx/substrate interface regime has a noticeable influence on the threshold voltage. S. Burignat [44] showed this effect using double derivative method to extract V_t depending on V_B on both Ultra-Thin Body (UTB) and UTBB devices. Result is shown in Figure 2-9b. From this figure we see that V_t is almost constant when the substrate is depleted. In this case the back bias sweep is partially compensated by the variation of the potential drop in the substrate depletion region, flattening the V_t-V_B curve.

2.1.3 Inversion charge summary

After deriving Q_i and V_t for the bulk case, we have then adapted the approach for the case of FD-SOI. We showed in all cases that we can model Q_i according to the equation below in strong inversion:

$$Q_i = -C_{ox}(V_G - V_t) \tag{38}$$

We have investigate the impact of using intrinsic or doped channel, the effect of BOx and channel thickness as well as the substrate depletion for the case of UTBB devices. The V_t-V_B behavior of FD-SOI UTBB structure with intrinsic channel has five different regimes with two that are not physically reachable (only mathematical derivation and numerical simulations can assess it). These are the case

of front inversion with back accumulation and front accumulation with back inversion. Regimes that are physically feasible are depicted by S. Burignat [44] equivalent capacitance model in Figure 2-10:

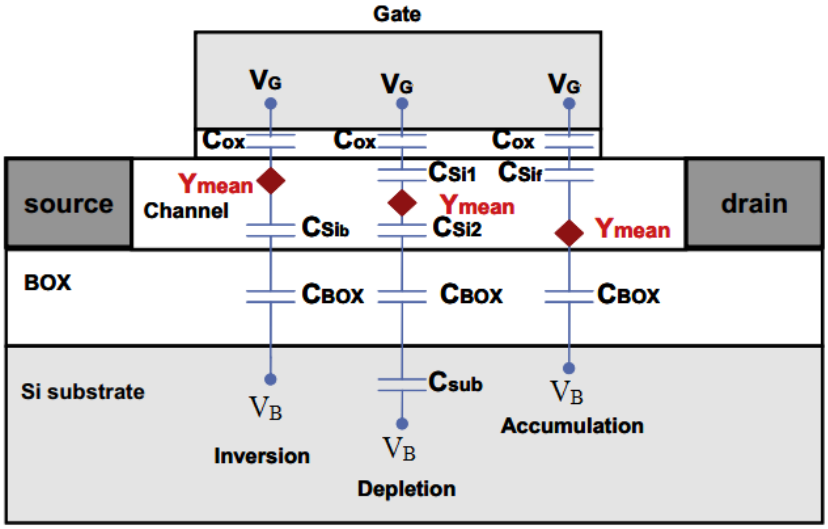


Figure 2-10: Schematic presenting the effective capacitances for the three main substrate regimes.[44]

In this figure, the three regimes are distinguished by the position of the inversion layer in the silicon channel. The charge centroïd moves from the front to back side when V_B goes from negative to positive values. This is modeled by considering the charge centroïd Y_{mean} in the channel and the silicon capacitance film on both sides of it (C_{sib} and C_{sif}). Thereafter we summarize V_t equation depending on these regimes.

- Front channel, back interface depletion, substrate inversion.

In this case the inversion layer is confined at the front interface and the back interface is depleted. This situation is depicted in Figure 2-10 (left side) and V_t is deduced from (35) where $2\phi_b$ is replaced by φ_{th} and Q_d is neglected:

$$V_t = V_{FBF} + \phi_{th} - (V_B - V_{FBB} - \phi_{th}) \cdot \frac{C_{si}C_{box}}{C_{ox}(C_{si} + C_{box})} \tag{39}$$

- Channel in the middle of the silicon film, back and front interface depletion, substrate depletion

When the substrate is depleted, C_{ox} and C_{box} are replaced by equivalent capacitance. The equivalent BOx capacitance yields:

$$C_{box_{eq}} = \frac{C_{box}C_{sub}}{C_{box} + C_{sub}} \tag{40}$$

and the equivalent oxide capacitance yields:

$$C_{ox_{eq}} = \frac{C_{ox}C_{si1}}{C_{si1} + C_{ox}} \tag{41}$$

Where $C_{si1} = \varepsilon_{si}/Y_{mean}$. C_{sub} is the substrate depletion. V_t equation is deduced from (39) where C_{ox} and C_{BOx} are replaced by their equivalent formulation and $C_{si} = \varepsilon_{si}/(Tsi - Y_{mean})$:

$$V_t = V_{FBF} + \phi_{th} - (V_B - V_{FBB} - \phi_{th}) \cdot \frac{C_{si}C_{box}C_{sub}(C_{ox} + C_{si1})}{C_{ox}C_{si1}(C_{si}C_{BOx} + C_{si}C_{sub} + C_{box}C_{sub})} \tag{42}$$

- Back channel, front interface depletion, substrate accumulation

In this case the inversion layer is confined at the back interface and the front interface is depleted. This situation is depicted in Figure 2-10 (right side) and V_t is deduced from (33) where $2\phi_b$ is replaced by φ_{th} and Q_d is neglected:

$$V_t = V_{FBF} + \phi_{th} - (V_B - V_{FBB} - \phi_{th})\frac{C_{box}(C_{si} + C_{ox})}{C_{ox}C_{si}} \tag{43}$$

To conclude, expected dependence of threshold voltage on back bias for UTBB devices is shown in Figure 2-11:

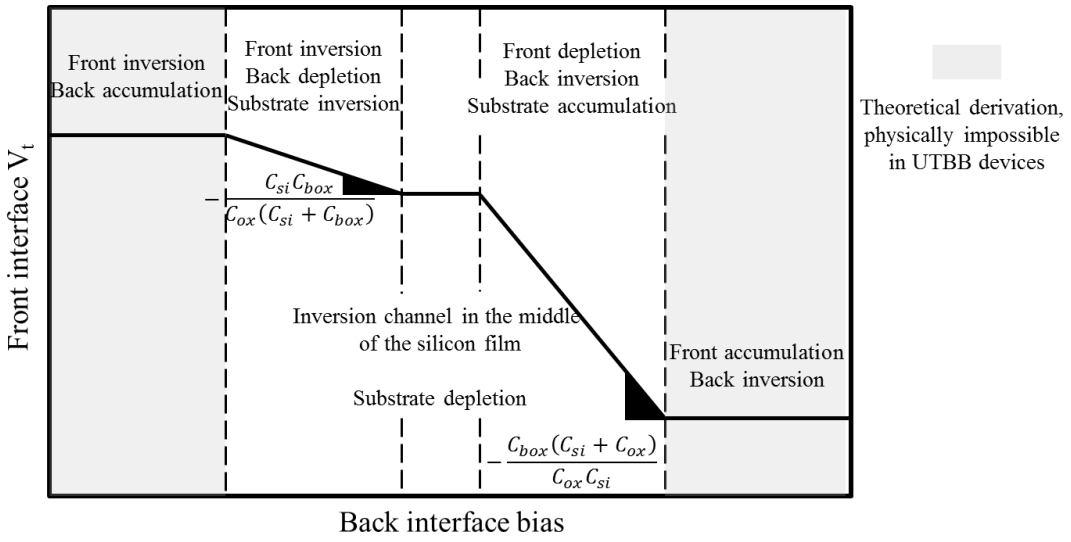


Figure 2-11: Theoretical Vt-VB curve with the different channel regimes for UTBB devices.

This figure represents the three physically common regimes (white areas) and the two physically unreachable regimes (shaded areas).

2.2 Channel carrier mobility

In previous section we have investigated the inversion layer carrier concentration in both bulk and FD-SOI transistors. It will be the basis for drain current formulation. But before, this section is devoted to the effective mobility experienced by inversion layer carriers. We will go through the main physical phenomenon that limit the mobility and then propose a compact model to address it, which will be used for the drain current formulation in section 2.3 and 2.4.

2.2.1 The effective mobility

Carrier mobility is limited by scattering phenomenon of different natures. Each phenomenon is characterized by a scattering time τ and can be written under the form: $\mu = e.\tau/m^*$ where m^* is the effective mass. Resulting effective mobility of all these scattering mechanisms is expressed following the Mathiessen rule:

$$\frac{1}{\mu_{total}} = \sum_i \frac{1}{\mu_i} \tag{44}$$

The three main scattering mechanisms are phonon, Coulomb scattering and surface roughness limited mobility. The limiting scattering mechanism is the one that has the smallest intrinsic mobility and each of them depends on the effective transversal field and/or on the inversion charge density. This formulation leads to the well-known universal mobility as illustrated in Figure 2-12:

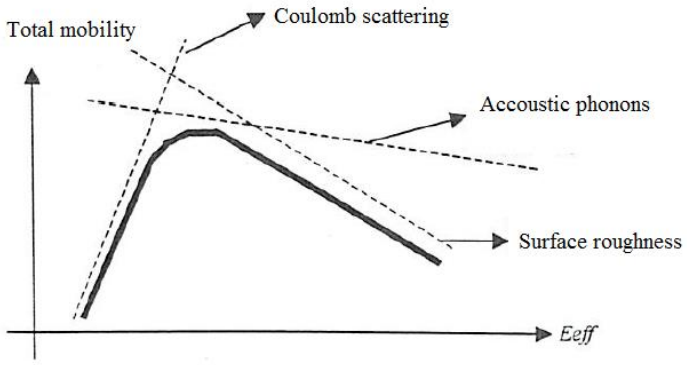


Figure 2-12: Universal mobility depending on effective field strength.

Phonon limited mobility has the following formulation [45]:

$$\mu_{ph} = A.E_{eff}^{-0.3}.T^{-\alpha} \tag{45}$$

where T is the temperature and E_{eff} the effective transversal electric field and $-1 < \alpha < -1.75$. Surface roughness limited mobility has the following formulation [45]:

$$\mu_{sr} = B.E_{eff}^{-\beta} \tag{46}$$

A and B are experimental fitting parameters and $-2 < \beta < -2.6$. Finally Coulomb limited mobility is usually experimentally extracted from measured effective mobility and universal mobility that accounts for surface roughness and phonon scattering, following [45]:

$$\mu_{coul}^{-1} = \mu_{eff}^{-1} - \mu_{universal}^{-1} \tag{47}$$

There are two types of Coulomb scattering to be considered: scattering with ionized dopants located in the channel and scattering with charges located in the gate stack or at the interface between materials, also called Remote Coulomb Scattering (RCS). Since we are working with undoped SOI MOSFET, we expect the first mechanism to be of little importance. However UTBB devices involve thin silicon

channel and thus increase the effect of RCS limited mobility as shown by D. Esseni [46] and C. Fenouillet-Beranger [47]. In these publications, the carrier mobility is almost only impacted by remote Coulomb scattering. A compact formulation for RCS limited mobility has been proposed by G. Hiblot [48] that is a trade-off between J. Koga [49] and F. Boeuf [46] approaches:

$$\frac{\mu_{C,it}}{\mu^0_{C,it}} = \frac{\sqrt{Q_i} + 0.1\sqrt{|q|D_{it}}}{0.1\sqrt{|q|D_{it}}} \tag{48}$$

where $\mu_{C,it}$ is the RCS mobility due to interface charge, D_{it} is the surface density of interface charges, and $\mu^0_{C,it}$ is the unscreened interface traps mobility.

To conclude, the mobility can be expressed as a function of the effective field following Matthiessen rule and neglecting Coulomb limited mobility:

$$\mu = \frac{1}{A^{-1}.E_{eff}^{0.3}.T^{1.75} + B^{-1}.E_{eff}^{2.6} + C^{-1}.Q_i^{-0.5}} \tag{49}$$

where $C = \frac{5450.10^{11}}{D_{it}.\sqrt{|q|D_{it}}} cm^2/V/s$ is a fitting parameter.

2.2.2 Mobility compact modeling

As we are considering the FD-SOI transistors, effective field is expressed as [50]:

$$E_{eff} = \frac{Q_d + \eta Q_i}{\varepsilon_{si}} + E_{sb} \tag{50}$$

Q_d can be neglected considering undoped channel. η equals 1/2 for electrons and 1/3 for holes [51]. Considering $\frac{dE_{eff}}{dQ_i}$ constant, a second order expansion of 1/μ as a function of Q_i yields:

$$\frac{1}{\mu} = D + E.Q_i + F.Q_i^2 \tag{51}$$

where D, E and F are factors that depend on E_{sb} η and ε_{si}. Then knowing that $Q_i \propto (V_G - V_t)$, μ yields:

$$\mu = \frac{\mu_0}{1 + \theta_1(V_G - V_t) + \theta_2(V_G - V_t)^2} \tag{52}$$

where μ_0 is the low field mobility and θ_1 and θ_2 account for effective field dependent mobility correction. These parameters depend on C_{ox}, E_{sb}, η and ε_{si}. This formulation will be kept for later model since it is compact and handy. However a rigorous derivation of this effective mobility can be performes analytically. Replacing Q_i by (38), E_{sb} by (22), the transverse electric field reduces to:

$$E_{eff} = \frac{\eta Cox}{\epsilon_{si}}(V_G - V_t) - \frac{C_{box}}{\varepsilon_{si}}(V_B - V_{FBB} + \varphi_{sb}) \tag{53}$$

Then φ_{sb} is replaced by its expression derived from (26) at threshold (Q_i=0 and Q_d is neglected):

$$\varphi_{sb} = \frac{C_{box}}{C_{si}+C_{box}}\left(V_B - V_{FBB} + \frac{C_{si}}{C_{box}}\varphi_{th}\right) \tag{54}$$

V_B can be expressed as a function of V_t following (39) and (43).

$$V_B = \frac{1}{\alpha}(V_{FBF} - V_t + \varphi_{th}) + V_{FBB} + \varphi_{th} \tag{55}$$

Where $\alpha = \frac{C_{si}C_{box}}{C_{ox}(C_{si}+C_{box})}$ when the channel is at the front interface and the back interface is depleted and $\alpha = \frac{C_{box}(C_{si}+C_{ox})}{C_{ox}C_{si}}$ when the back interface is inverted and the front is depleted. Finally inserting (55) and (54) into (53), E_{eff} reduces to:

$$E_{eff} = \frac{\eta C_{ox}}{\epsilon_{si}}\left(V_G + V_t\left(\frac{r}{\eta} - 1\right)\right) + s \tag{56}$$

where r depends on capacitances and the back interface regime. r formulas are written below depending on the back interface regime:

(back interface depleted)
$$r = \frac{2C_{BOx}}{C_{si}} + 1 \tag{57}$$

(back interface inverted)
$$r = \frac{C_{si}(C_{si}+2C_{box})}{(C_{si}+C_{ox})(C_{si}+C_{box})} \tag{58}$$

and s depends on flatband voltages, φ_{th} and the back channel regime:

(back interface depleted)
$$s = \left(C_{si} + 2C_{box} + \frac{2C_{si}C_{box}}{C_{ox}}\right)\varphi_{th} + (C_{si}+2C_{box})V_{FBF} \tag{59}$$

(back interface inverted)
$$s = \left(\frac{2C_{box}^2 + 4C_{ox}C_{box}}{C_{ox}} + \frac{2C_{box}^2}{C_{si}} + \frac{C_{si}(C_{ox}+2C_{box})}{C_{ox}}\right)\varphi_{th} + (C_{si} + 2C_{box})V_{FBF} \tag{60}$$

To conclude, the effective mobility depends on E_{eff} that is proportional to $V_G + \left(\frac{r}{\eta} - 1\right)V_t$ as shown in (56) and Q_i that is proportional to $V_G - V_t$. This conclusion has been verified by simulation [52] where we can see a universal behavior of mobility depending on Q_i (on E_{eff}) only where Coulomb scattering dominates (is negligible). So, the compact model for μ in (52) is a rough approximation and considering E_{eff} expression (56) and the mobility expression (49), it can be shown that θ_1, θ_2 and μ_0 fitting parameters in expression (52) depend on C_{ox}, C_{si}, C_{box}, V_{FBF} and V_t. Thus θ_1 and θ_2 have neither universal behavior nor physical meaning.

2.2.3 Mobility degradation for short channel devices

Mobility degradation of short channel devices has been widely investigated in literature. Multiple explanations have been proposed for that phenomenon. First, Ghibaudo [53], Barral [54], Pappas [55] and Guarnay [56] investigated the effect of ballistic transport on the effective mobility in linear regime. Carriers experience ballistic transport if no scattering mechanism affects their transport. It occurs in very short channel devices. Indeed, considering that scattering events occur periodically, if time needed for the carriers to cross the channel is smaller or comparable to the relaxation time of scattering mechanisms, then carriers can experience no scattering. Ghibaudo [53] addresses ballistic transport using quantum mechanics. His ballistic drain formulation yields:

$$I_{D_{bal}} = \frac{W}{2} C_{ox} v_{inj} (V_G - V_t) \cdot \frac{qV_D}{kT} \tag{61}$$

where

$$v_{inj} = \sqrt{2k_B T/(\pi m_t)} \tag{62}$$

v_{inj} is the injection velocity. m_t is the transverse electron mass ($m_t = 0.19m_0$ where m_0 is the free electron mass), k_B is the Boltzmann constant and T the temperature. In this case, equivalent mobility formulation yields:

$$\mu_{eq} = \frac{q v_{inj} L}{2k_B T} \tag{63}$$

where L is the channel length. Equation (63) shows that μ_{eq} is proportional to L, thus the shorter is the channel, the lower is the apparent mobility. However Ghibaudo [53], Fleury [57], Barral[54], Pappas[55] and Shin [58] showed that the proportion of ballistic transport is rather low even for the shortest devices and ballistic mobility cannot explain the entire apparent mobility degradation. Finally recent Monte-Carlo studies [56], [59] and [60] suggested that ballistic transport contribution could be underestimated depending on the extraction method used for the backscattering coefficient extraction from mobility measurements.

The second explanation for mobility degradation is saturation velocity. It limits the mobility at high lateral field (thus more important for short channel devices). Carrier may reach the saturation velocity v_{sat} if the lateral field is high enough ($E_{lat}>E_{crit}=10^4$V/cm). Indeed, when the carrier reaches v_{sat} it has sufficient energy to generate a phonon. The energy required to create the phonon is taken from the carrier, reducing its velocity. Average carriers velocity is then clamped to v_{sat}. Experimental saturation velocity measurements have first been done by Ryder [61]. Further investigations to reach v_{sat} have been done on P-N junctions by Norris and Gibbons [62], Duh, Moll [63] and Rodriguez, Ruegg and Nicolet [64]. Average values are $v_{sat}=10^7$cm/s for electrons and 6.10^6 cm/s for holes in silicon crystal.

However this phenomenon appears at equilibrium state. Indeed, the carrier velocity is only constant when averaged over a time much longer than the scattering time (τ). In short devices, the time required for an electron to cross the channel is comparable to τ. Thus the carrier can cross the channel in a transient state enabling a velocity larger than v_{sat}, this phenomenon is called velocity overshoot. Ruch [65] used Monte Carlo simulations to demonstrate this effect on GaAs transistors. Recent investigations done by Kim et al. [66] showed experimental observations of carrier reaching velocity overshoot.

Lundstrom [67] suggested that drain saturation current can be modeled easily using barrier scattering theory. Moreover he suggests that velocity overshoot should not be the prominent effect in ultimately scaled MOSFET since carrier are cold injected at the source and may only overshoot v_{sat} if the channel is sufficiently long. His drain current formulation yields:

$$I_{D\_Vsat} = \frac{WC_{ox}}{\frac{1}{v_{inj}} + \frac{1}{\mu_{eff}\epsilon(0^+)}}(V_G - V_t) \tag{64}$$

where $\epsilon(0^+)$ is the maximum potential barrier height at the virtual source and $v_{inj} = \sqrt{2k_BT/(\pi m_t)}$. His interpretation is close to Natori's one in (61) that addresses ballistic transport. Thus distinguishing ballistic transport and velocity saturation is not easy. The relevance of these approaches is discussed by Yang et al. [68]. They showed some contradictions with experiments in particular about temperature dependencies. To overcome this, they investigated which saturation effect intervenes depending on the device geometries and the operating conditions and proposed a unified model for saturation velocity and ballistic transport, mixing Lundstrom [67] (64) and Natori's one [69] (61) that yields:

$$I_{D\_Vsat} = \frac{WC_{ox}}{\frac{3\pi m_t\sqrt{q\pi M_v}}{4h\sqrt{C_{ox}(V_G - V_t)}} + \frac{1}{\mu_{eff}\epsilon(0^+)}}(V_G - V_t) \tag{65}$$

Where M_v is the product of the lowest valley degeneracy and the reciprocal of the fraction of the carrier population in the lowest energy level. In their study it is shown that determining whether the drain current is v_{inj}, v_{sat}, velocity overshoot or pinch off limited is a tricky task and requires comprehensive characterization including temperature dependence.

In order to meet our goal we only need phenomenological approach. Thus saturation velocity effect is accounted for by introducing it through the mobility such as:

$$\mu_{short} = \frac{1}{\frac{1}{\mu} + \frac{E_{lat}}{v^*}} \tag{66}$$

where μ is the mobility as formulated in (52) and E_{lat} is the lateral field. This approach has been reported in many compact models as in [70]-[72]. As mentioned previously, ballistic and velocity saturation limited currents have the same form, thus v^* can account for both v_{inj} and v_{sat} as suggested by Fleury [57].

Another effect involves extra scattering mechanisms induced by neutral or charged defects at the S/D channel junction. These defects are induced by S/D dopants implantation. For long channel devices, most of the channel is defect free but for short channel devices, S/D junction is a significant part of the transport region thus apparent mobility is driven by neutral defects scattering mechanism. Ghibaudo [53] used temperature dependent mobility extraction on FD-SOI, double gate and gate-all-around MOSFET as well as FinFET transistors. This approach enables distinguishing the contribution of neutral defects scattering from other scattering mechanisms, revealing its dominant effect. Pham-Nguyen [73] confirmed these results using different gate stacks, Cassé [74] and Chaisantikulwat [75], Shin [76][58] confirmed it using magnetoresistance measurements. Finally, Barral [54] and Pappas [55] confirmed it by extracting the backscattering coefficient from mobility measurements. It should be noted that recent extraction done on 14 nm FD-SOI MOSFETs with in situ doped raised source

drain technology showed the same gate length mobility roll-down [77][78]. However, in situ doping is expected to reduce the formation of neutral defect since there is no implantation during the process. Thus the mobility reduction cannot be explained only by the presence of neutral defects.

As a conclusion, quantifying precisely the contribution of each physical mechanism to the channel length mobility roll down is a tricky task. However Shin [58] showed that the apparent mobility degradation measured using magnetoresistance can be modeled as:

$$\mu_{short} = \frac{\mu}{1 + \frac{L_c}{L}} \tag{67}$$

Where μ is expressed as in (52) and L_c is the critical length at which μ_{short} is half the long channel mobility μ. This empirical model fit well the apparent mobility. Combining (52), (67) and (66) yields the effective mobility accounting for remote Coulomb and phonon scattering, surface roughness, velocity saturation, ballistic transport and neutral defects:

$$\mu_{eff} = \frac{\mu_0}{\left(1 + \theta_1(V_G - V_t) + \theta_2(V_G - V_t)^2 + \frac{E_{lat} \cdot \mu_0}{v^*}\right) \cdot \left(1 + \frac{L_c}{L}\right)} \tag{68}$$

2.3 Linear regime model

In this section we derive the linear drain current equation based on previously investigated quantities like V_t, Q_i and μ_{eff}. The transistor structure is depicted in Figure 2-13.

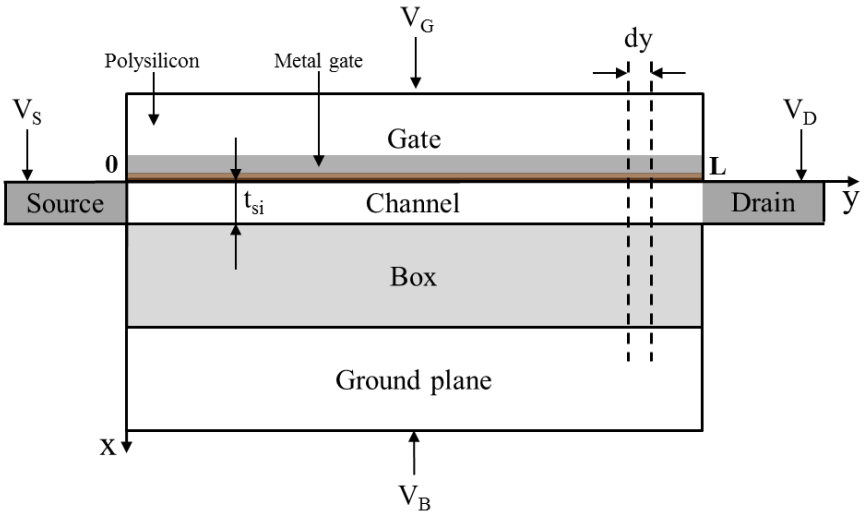


Figure 2-13: Basic element of FD-SOI MOSFET architecture.

In our case the transistor is in strong inversion regime ($V_G > V_t$). An inversion layer is created in SOI. Considering the case of nMOS where $V_S = 0$, if $V_D > 0$, electrons start to drift toward the drain end with a mobility as described in §2.2. A current flows through the MOS. The total current is constant along the y direction. Then at a point y in the channel, the inversion charge is $Q_i(y)$ and the mobility is $\mu_{eff}(y)$. The potential drop across an infinitesimal section dy induces a current expressed as:

$$I_{Dlin} = Q_i \cdot \mu_{eff} \cdot E_{lat} \tag{69}$$

with $E_{lat} = dV_c/dy$ being the electric field in the y direction. The current is conserved along the channel, thus, integrating the expression along the width direction, the drain current yields:

$$I_{Dlin} = -W \cdot Q_i(y) \cdot \mu_{eff}(y) \cdot \frac{dV_c}{dy} \tag{70}$$

where W is the width of the transistor, μ_{eff} is the effective carrier mobility reported in (68) and Q_i is the inversion layer carrier density and dV_c is the potential drop across the channel section, V_c being the quasi Fermi potential along the channel. If we assume that Q_i and μ_{eff} do not explicitly depend on y but only on V_c, then y and V_c variables can be separated following:

$$I_{Dlin} dy = -W \cdot Q_i(V_c) \cdot \mu_{eff}(V_c) \cdot dV_c \tag{71}$$

This assumption is valid if E_{lat} is constant along the channel, that is true in linear regime. The drain current is then obtained by integrating (71) from the source to the drain:

$$\int_0^L I_{Dlin}\, dy = -W \int_{Vs}^{Vd} \mu_{eff}(V_c) Q_i(V_c) dVc \tag{72}$$

Equation (72) shows that I_D calculation requires the inversion carrier concentration Q_i. To use previously derived equation for Q_i in §2.1, a variable change should be operated. Indeed V_G and φ_{sf} biases were referred to a hypothetical unbiased neutral body that was $E_f/q=V_B$ in the case of MOS capacitor. Here $E_f/q=V_c$ is no longer equal to the back bias but is driven by the source-drain bias and goes linearly (for the case of strong inversion) from V_S up to V_D. From now on, we will use V_S as the reference potential. Thus φ_{sf} shall become $\varphi_{sf} + V_c$ and V_G become $V_{GS} - V_C$. Depending on whether we consider or not θ_1 and θ_2 parameters, I_{Dlin} yields the formulas listed in Table 2-1.

$\theta_1 = \theta_2 = 0$	$I_{Dlin} = \frac{W.C_{ox}.\mu_0.V_D}{(L+L_c)A}\left(V_G - V_t - \frac{V_D}{2}\right)$	(73)
$\theta_2 = 0$ $\theta_1 \neq 0$	$I_{Dlin} = \frac{W.C_{ox}.\mu_0}{(L+L_c).\theta_1^2}\left(\ln\left(1 - \frac{V_D\theta_1}{A + \theta_1(V_G - V_t)}\right)A + V_{DS}.\theta_1\right)$	(74)
$\theta_1 = 0$ $\theta_2 \neq 0$	$I_{Dlin} = \frac{W.C_{ox}.\mu_0}{2.(L+L_c).\theta_2}\ln\left(\frac{\theta_2(V_G - V_t - V_D)^2 + A}{\theta_2(V_G - V_t)^2 + A}\right)$	(75)
$\theta_1 \neq 0$ $\theta_2 \neq 0$	$Id_{Dlin} = I'_{Dlin}(Vd) - I'_{Dlin}(0)$ $I'_{Dlin}(u) = \frac{W.C_{ox}.\mu_0}{L+L_c}.\left((\ln(a) - \ln(b))\frac{\theta_1}{2\theta_2}\sqrt{\frac{1}{\theta_1^2 - 4A\theta_2}} - \frac{1}{2\theta_2}(\ln(a) + \ln(b))\right)$ $a = V_t - V_G + u - \frac{\theta_1}{2\theta_2} + \left(\frac{\theta_1^2}{2\theta_2} - 2A\right)\sqrt{-\frac{1}{4A.\theta_2 - \theta_1^2}}$ $b = V_t - V_G + u - \frac{\theta_1}{2\theta_2} - \left(\frac{\theta_1^2}{2\theta_2} - 2A\right)\sqrt{-\frac{1}{4A.\theta_2 - \theta_1^2}}$	(76)
	In above formulas $A = 1 + \frac{V_{DS}\mu_0}{Lv^*}$	(77)

Table 2-1: I_{Dlin} formulation for long channel devices with and without θ1 and θ2 parameters.

Notice that I_{Dlin} formulation when θ_1 and θ_2 are not null, yields only real values if $4A\theta_2 < \theta_1^2$.

In order to simplify these expressions, μ_{eff} can be assumed independent of the position along the channel and replaced by its average value in the channel [32]. Initially, μ_{eff} is expressed as in (68), including the variable change for V_G:

$$\mu_{eff} = \frac{\mu_0}{\left(1 + \theta_1(V_{GS} - V_t - V_c) + \theta_2(V_{GS} - V_t - V_c)^2 + \frac{V_{DS} \cdot \mu_0}{v^* L}\right) \cdot \left(1 + \frac{L_c}{L}\right)} \tag{78}$$

A good approximation of the average mobility along the channel is obtained by replacing V_c by $V_{DS}/2$. It can then be taken out from the integral (72) and the drain current yields:

$$I_{Dlin} = \frac{\mu_{eff} W}{L} \int_{Vs}^{Vd} C_{ox}(V_{GS} - V_t - V_c) dV_c \tag{79}$$

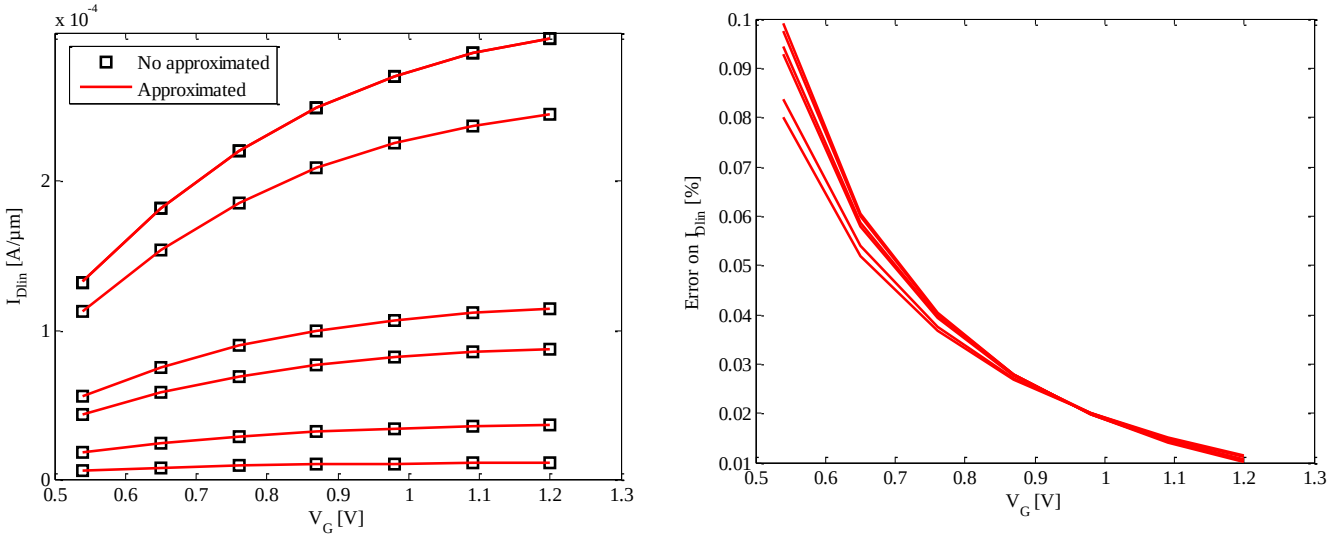
Integrating (79) yields:

$$I_{Dlin} = \frac{\mu_{eff} W}{L} C_{ox} \left(V_{GS} - V_t - \frac{V_{DS}}{2}\right) V_{DS} \tag{80}$$

In this formula the effective mobility has been replaced by (68) where $V_G - V_t$ has been replaced by $V_{GS} - V_t - \frac{V_{DS}}{2}$. The potential drop $\frac{V_{DS}}{2}$ is the average of V_c along the channel. The lateral electric field E_{lat} in (68) has been replaced by its average $E_{lat}=V_{DS}/L$.

This equation is valid in strong inversion regime ($V_{GS} > V_t$) and a second order high field mobility reduction formulation is assumed here. This equation does not take into account the effect of access resistance. This restriction will be discussed in §2.5.

We shall see later for the parameters extraction that the linearized formulation of I_{Dlin} (80) is more convenient and will be preferred over (76). The error between I_{Dlin} approximated in (80) and not approximated in (76) is shown in Figure 2-14 against V_G with different channel lengths. We see that the error in the right plot is low (below 0.1%) and depends slightly on L.



(a) (b)

Figure 2-14: I_{Dlin} against V_G using the approximated (72) and the not approximated (76) formula for $L \in [1; 0.02]$ µm. For the calculation, the following parameters have been used: µ=200 cm²/V/s, C_{ox}=3*10⁻⁶ F/cm², Vt = 0.3 V, $\theta_1 = 1\,V^{-1}$, $\theta_2 = 1V^{-2}$. V_{DS}=0.05V.

Notice that saturation velocity effect is driven by the lateral field. Thus its effect will only be significant at high V_{DS}, in saturation regime. Hence, v^* contribution is neglected in linear regime. To conclude, the general formulation for the linear drain current we will use is:

$$I_{Dlin} = \frac{W}{L + L_c} \frac{\mu_0 C_{ox}}{1 + \theta_1 \left(V_{GS} - V_t - \frac{V_{DS}}{2}\right) + \theta_2 \left(V_{GS} - V_t - \frac{V_{DS}}{2}\right)^2} \left(V_{GS} - V_t - \frac{V_{DS}}{2}\right) V_{DS} \quad (81)$$

This formulation is derived using simplification but these are fully justified, as shown in Figure 2-14.

2.4 Saturation regime model

In this section we will introduce the saturation mechanism that occurs at high V_D. We will show that in the case of long channel devices, the saturation is caused by pinch-off phenomenon. From the linear drain current model we will derive a saturation drain voltage V_{Dsat} and deduce the saturation drain current.

2.4.1 Effect of high drain voltage: pinch off saturation

In the case of long channel transistors, the saturation mechanism that occurs is called pinch-off [79]. This phenomenon can be explained with the help of Figure 2-15. In linear ($V_{DS} << V_{GS} - V_t$) and strong inversion ($V_{GS} > V_t$) regimes, the inversion layer is uniform in the channel. The channel acts as a resistor and the drain current verifies equation (81). Then, as the drain voltage rises, the depletion area around the drain increases, reducing Q_{inv} at the drain end. When $V_{DS} = V_{Dsat}$ the inversion charge is null at the channel drain junction. This is the pinch off point. Then if V_{DS} is even more increased, the inversion layer continues to shrink and the pinch off point drifts from the drain channel junction toward the source side. Any increase in drain voltage is compensated by a voltage drop across the depletion region at the drain end and not by an increase of the current. This is why the current saturates beyond the pinch off point.

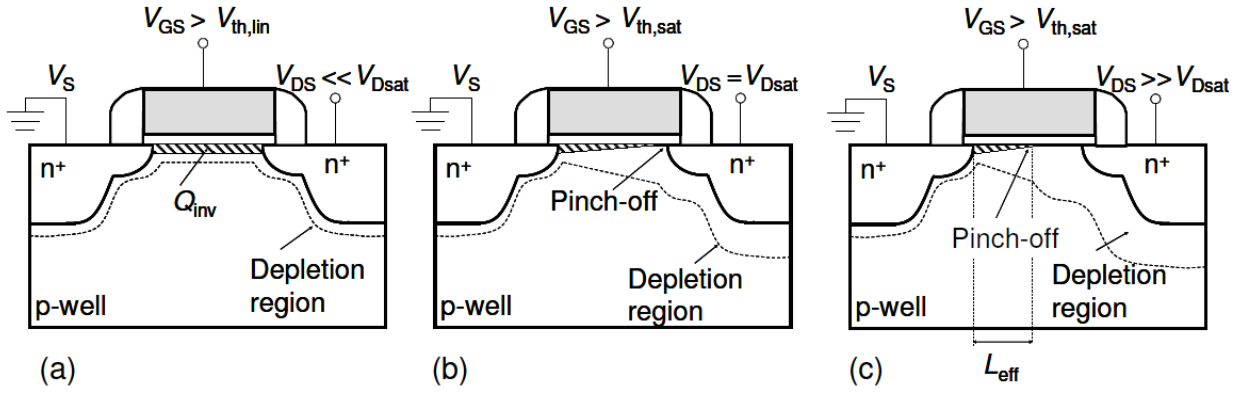


Figure 2-15: NMOS transistor operating in (a) the linear model, (b) the onset of saturation, and (c) beyond saturation where the effective channel length L_{eff} is reduced. $V_{th,lin}$ and $V_{th,sat}$ are the linear threshold voltage and the saturation threshold voltage, respectively. Q_{inv} is the inversion charge. [68]

Mathematical formulation of the saturation drain current is derived from the linear one (72) where v^* is accounted in the effective mobility since V_{DS} and E_{lat} are high in this case. I_{Dlin} reaches a maximum at $V_D = V_{Dsat}$ as shown in Figure 2-16. Thus, a common practice to derive V_{Dsat} is to calculate $G_{DS} = \frac{dI_{Dlin}}{dV_D}$ and find V_D where $G_{DS} = 0$. Using I_{Dlin} equation (76), no analytical formula for V_{Dsat} and

the saturation drain current can be found. A first approximation consists in replacing E_{lat} by $\frac{V_{DS}}{L}$. Then I_{Dsat} formulation can be found by taking I_{Dlin} formula in Table 2-1 and replacing V_{DS} by V_{Dsat} derived from $G_{DS} = 0$. This yield:

$$I_{Dsat} = -\frac{W}{L}\int_{Vs}^{V_{Dsat}} \mu_{eff}(V_c)Q_i(V_c)dVc \tag{82}$$

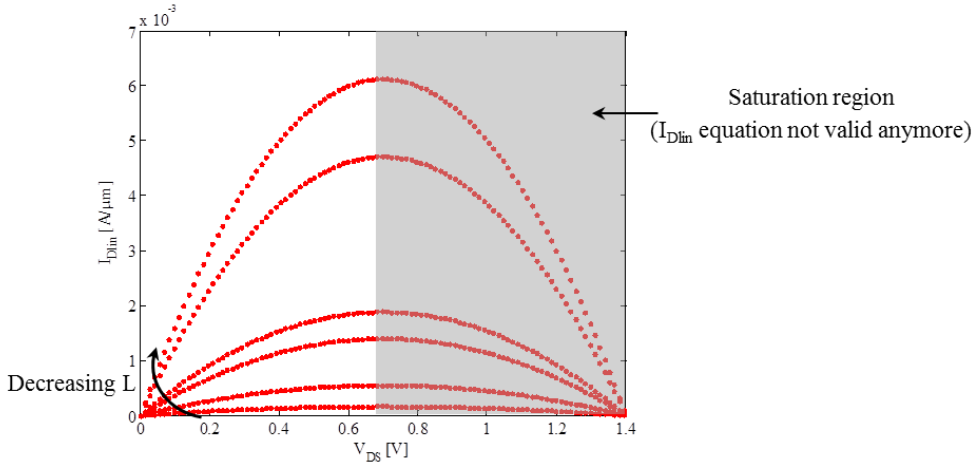


Figure 2-16: Width normalized I_D-V_D curves from equation (81) with $V_{GS} = 1V$. L=[0.024, 0.031, 0.078, 0.105, 0.267, 0.897] µm. $V_t = 0.3$ V, $\mu_0 = 200\frac{cm}{V.s}$, $\theta_1 = \theta_2 = 0$ V^{-1}. Effects of v^* and L_c are neglected here.

However V_{Dsat} formulation is not analytical making I_{Dsat} time consuming to compute, thus we shall find an analytical approximation. This is done following the same approach but starting with the simplified I_{Dlin} expression as expressed in equation (81) instead of (76). If $\theta_1 = \theta_2 = 0$ and v^* and L_c effects are neglected (long transistors), $V_{Dsat} = V_{GS} - V_t$ and the saturation drain current I_{Dsat} yields:

$$I_{Dsat} = \frac{\mu_0 W}{2.L} C_{ox}(V_G - V_{tsat})^2 \tag{83}$$

For the general case, V_{Dsat} yields:

$$V_{Dsat} = 2\frac{u - \sqrt{u.\left(1 + 2.\frac{\mu_0}{v^*.L}(V_G - V_{tsat})\right)}}{\theta_1 - \frac{2\mu_0}{L.v^*} + (V_G - V_{tsat})\theta_2} \tag{84}$$

where $u = 1 + \theta_1(V_G - V_{tsat}) + \theta_2(V_G - V_{tsat})^2$ and V_{tsat} is the threshold voltage at the operating voltage $V_{DS} = V_{DD}$.

To conclude the general formulation for the saturation drain current yields:

$$Id_{sat} = \frac{W}{L + Lc}\frac{\mu_0 C_{ox}}{1 + \theta_1\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right) + \theta_2\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)^2 + \frac{V_{Dsat}.\mu_0}{Lv^*}}\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)V_{Dsat} \tag{85}$$

This equation is valid in strong inversion regime ($V_{GS} > V_t$) with no access resistance. This restriction will be discussed in §2.5.

Notice that if saturation velocity, ballistic transport and neutral defects effects are small (long transistors), then Id_{sat} tends to have quadratic dependence with $V_{GS}-V_t$ as in (83) (pinch off saturation mechanism). However if saturation velocity limits the current, then Id_{sat} has a linear dependence with $V_{GS}-V_t$. This agrees with theory, following equations (64)-(65) for velocity saturation and ballistic transport and (83) for pinch off saturation. Hence, expression (85) is suited for saturation. L_c account for neutral defects and v^* accounts for velocity saturation and ballistic transport.

2.5 Effect of access resistance…

2.5.1 …in linear regime

In the previous paragraph, the drain current equation has been derived considering no effect of source drain (S/D) and contacts region. This assumption is valid for long channel devices (case where the access resistance is small compared to channel resistance). In this section we will investigate the impact of access resistance in short channel devices and include it into our model.

The first studies used to consider constant access resistance [80]-[82]. Including a constant access resistance in the drain formulation is done by substituting V_{GS} and V_{DS} respectively by $V_{GS}-R_S.I_D$ and $V_{DS}-(R_S+R_D).I_D$ where R_S and R_D are the S/D resistance. Considering equation (81) for the access resistance free linear drain current, I_{Dlin} with access resistance yields:

$$I_{Dlin}=\frac{\beta\left(V_{GS}-V_t-\frac{V_{DS}}{2}\right)V_{DS}}{\left(1+(\theta_1+R_{sd}\beta)\left(V_{GS}-V_t-\frac{V_{DS}}{2}\right)+\theta_2\left(V_{GS}-V_t-\frac{V_{DS}}{2}\right)^2\right)\left(1+\frac{Lc}{L}\right)} \quad (86)$$

With $\beta=\frac{W}{L}\mu_0 C_{ox}$ and $R_{sd}=R_S+R_D$. In terms of resistance, the width normalized total MOS resistance is:

$$R_{tot}=\frac{V_{DS}.W}{I_{Dlin}}=W\left(R_{sd}+\frac{1+\frac{Lc}{L}}{\beta}\left(\frac{1}{\left(V_{GS}-V_t-\frac{V_{DS}}{2}\right)}+\theta_1+\theta_2\left(V_{GS}-V_t-\frac{V_{DS}}{2}\right)\right)\right) \quad (87)$$

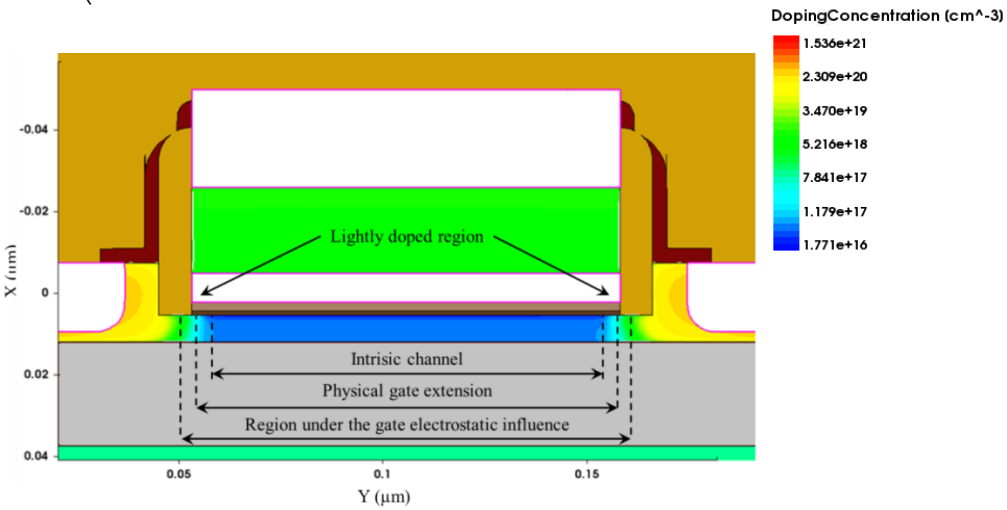


Figure 2-17: TCAD simulated MOS scheme with dopant concentration in silicon regions

Other studies have claimed that access resistance has a $1/(V_G - V_t)$ dependence [83]-[87][88]. The subtlety between these two hypotheses mostly lay in the definition of channel and access region. In order to demonstrate it, let's consider two cases. One case is where the channel is the region that encompasses all silicon regions where $N_a < N_{imax}$ where N_{imax} is the maximum inversion carrier density that could be induced by electric field. $N_{imax} \approx 10^{19} cm^{-3}$ [84]. The other case is where the channel is considered to be the intrinsic region of the channel that lies below the gate. These two situations are illustrated in Figure 2-17 and Figure 2-18.

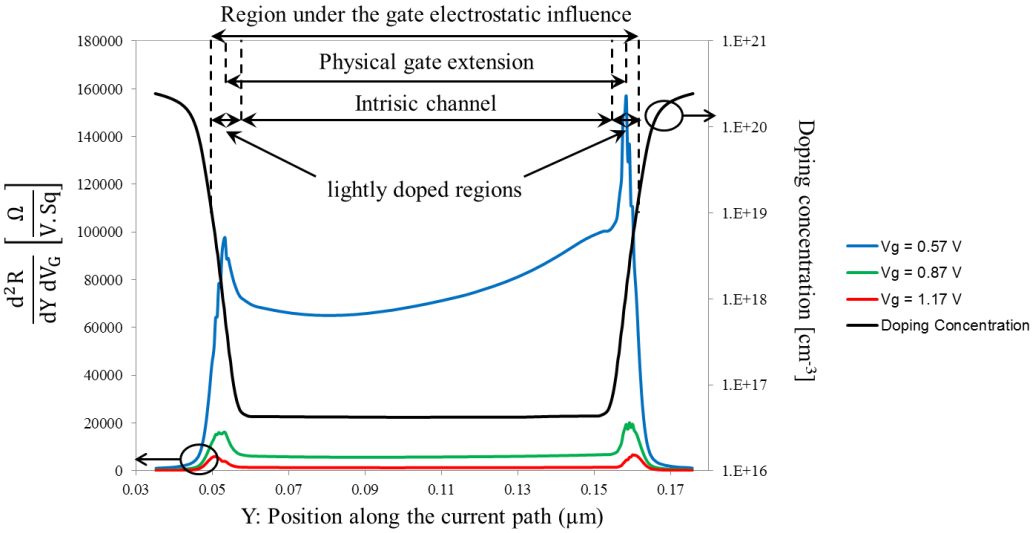


Figure 2-18: Local resistivity variation w.r.t V_{GS} against the position along the current path. Sq refers to the W/L normalization of the resistance.

Let's consider the first limit case. Here the access region carrier concentration can't be modulated by the gate bias and the access resistance is constant. In this case, the channel region can spread much beyond the physical gate length, encompassing lightly doped regions (LDR) and the average mobility in the channel will depend on the channel length due to Coulomb scattering at both ends of the channel.

In the second case the lightly doped regions are part of the access resistance where carriers will be accumulated as V_{GS} increases, reducing its resistivity. Thus in this case the access region is modulated by the gate voltage.

Figure 2-18 shows the local resistivity variation of the silicon at position y along the current path with respect to a small gate voltage variation (left axis). This curve shows where the silicon electrostatic is modulated by the gate potential. Obviously the silicon electrostatic is strongly dependent on the gate potential wherever the doping concentration is lower than 10^{19} atoms/cm^3. Doping concentration is shown in dark, referenced on the right vertical axis. This figure also shows that the physical gate extension, the intrinsic channel and the region under the gate electrostatic influence are not of the same size because the junctions are not perfectly sharp and aligned with the gate. Thus depending on the channel length that is considered, the access resistance will depend or not on the gate voltage. If we consider that L is smaller than the region under the gate electrostatic influence, then the access resistance is made of three contributions. The contact resistance, the highly doped source and drain resistance and the LDR resistance. Only the last contribution depends on V_G. Kim [89] and Taur [90] showed that the carriers in this region are accumulated and their concentration can be expressed as:

$$Q_{acc} = -C_{ox_{LDR}}(V_{GS} - V_{tLDR}) \tag{88}$$

V_{tLDR} is null if we are considering that carriers are ideally accumulated. However in the LDR, the doping concentration is not constant and depending on the position along the conduction path this V_{tLDR} will range from 0 (close to the highly doped region) up to V_t (close to the channel), as explained by Taur [90] and Sheu [88]. Nevertheless, this conclusion can be shaded. Indeed, this is true if we consider that the gate electrostatic field is uniforme along the conduction path up to the highy doped region. However in practical cases, lightly doped region can spread away from the gate (if the transistor is underlaped). In this configuration, LDR becomes hard to invert since it benefits from a weaker gate field. Consequently Vt_{LDR} can becomes higher than V_t. While Hu [83] considers $V_{tLDR} = V_t$ when extracting R_{SD} and L_{eff} depending on V_{GS} (what was justified considering the technology used at this time), Kim [91] showed that for more recent technologies, this simplification does not hold anymore. The consequence is that L_{eff} calculated using R_{tot}-L_{poly} curves gives unphysical effective channel length. This emphasizes the necessity to use $V_{tLDR} \neq V_t$ for the extraction. Thus the width normalized access resistance yields:

$$R_{sd} = R_0 + \frac{L_{LDR}}{\mu_{LDR} Cox_{LDR} (V_{GS} - V_{tLDR})} \tag{89}$$

with R_0 the width normalized contact resistance, L_{LDR} and μ_{LDR} the extension and the average carrier mobility of the LDR. Here V_{tLDR} is the LDR average threshold voltage. For our study we consider that L is the physical gate length, thus the width normalized total resistance yield:

$$R_{tot} = R_0 + \frac{\sigma}{V_{GS} - V_{tLDR}} + \frac{W\left(1 + \frac{Lc}{L}\right)}{\beta}\left(\frac{1}{\left(V_G - V_t - \frac{V_{DS}}{2}\right)} + \theta_1 + \theta_2\left(V_G - V_t - \frac{V_{DS}}{2}\right)\right) \tag{90}$$

To simplify the equation we used $\sigma = \frac{L_{LDR}}{\mu_{LDR} Cox_{LDR}}$ averaged over the lightly doped region. σ is small if the MOS is overlapped and large otherwise. This access resistance formulation has been reported and justified using TCAD simulation by Monsieur [86].

2.5.2 …in saturation regime

In order to derive I_{Dsat} rigorously, we consider I_{Dsat} formula **(82)**. Then R_S and R_D are accounted for by substituting V_{GS} by $V_{GS} - R_S . I_{Dsat}$ in I_{Dsat} equation **(82)**. I_{Dsat} is then found by solving this complex equation. Here $V_{DS} = V_{Dsat}$, thus we do not add access resistance through V_D but instead V_{GS} is substituted by $V_{GS} - R_S . I_{Dsat}$ in V_{Dsat} expression.

There is no analytical expression for I_{Dsat}, making it time consuming to compute. However it can be simplified by assuming equation (85) for the intrinsic saturation drain current I_{Dsat}. Then R_S effect is added using a first order expansion following:

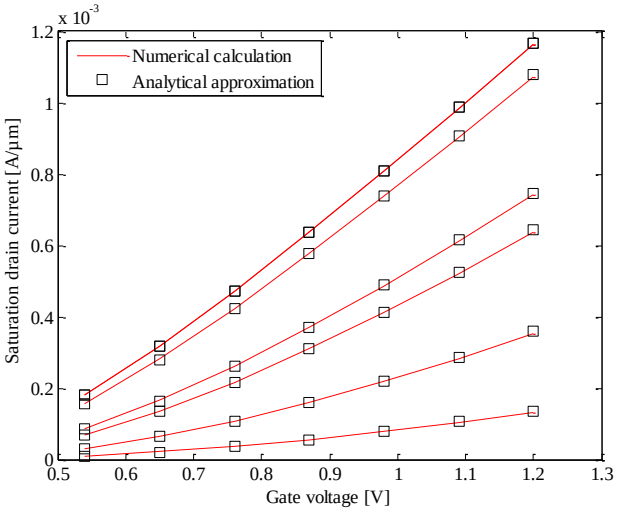
$$I_{Dsat} = \frac{I'_{Dsat}}{1 + G_m \cdot R_S} \tag{91}$$

Where I'_{Dsat} is the expression of intrinsic saturation drain current (see equation (85)) and $G_m = \frac{dId_{lin}}{dV_{GS}}$. Analytical formulation of G_m yields:

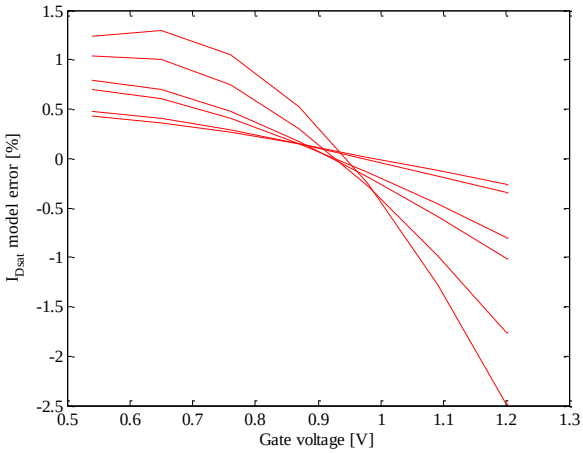
$$G_m = 4\frac{W}{L+L_c}\mu_{eff}C_{ox}V_{DS}.\left(A-\theta_2\left(V_{GS}-V_t-\frac{V_{DS}}{2}\right)^2\right) \quad (92)$$

where μ_{eff} is the effective mobility as described in **(78)** and $A = 1+\frac{V_{DS}.\mu_0}{Lv^*}$. V_{Dsat} is used in G_m expression instead of V_{DS} in order to get I_{Dsat} expression (91).

I_{Dsat} and V_{Dsat} values from (91) have been compared with the numerical solution of I_{Dsat} using formula (72) with access resistances. Results are shown in Figure 2-19. A good match is obtained using realistic model parameters. This comparison shows that approximating the effect of access resistance using (91) and considering a constant mobility along the channel are suited approximations.



(a)



(b)

Rs=Rd	90 Ω
μ_0	100 cm²/V/s
vsat	10^7 cm/s
V_D	1 V
W	1 µm
L	[1, 0.3, 0.12, 0.09, 0.038, 0.028] µm
Vt	0.3 V
θ_1	-0.5 V
θ_2	0.5 V

(c)

Figure 2-19: I_{Dsat} (a) calculated numerically using equation (85) with access resistance against simplified analytical expression (91). Model error in percentage is shown in (b). Model parameters used to compute I_{Dsat} are gathered in Table (c).

2.6 Conclusion

This chapter has provided an introduction to the basic equations that describe the drain current of MOSFET transistors. These equations will be used for further parameters extraction of our compact model.

Firstly, the MOS capacitance structure has been investigated to derive the inversion carrier concentration as well as the threshold voltage for the case of bulk devices. Then these equations have been adapted for the case of UTBB devices. The effects of channel doping concentration, ultra-thin channel and box on V_t and inversion charge density have been treated. A compact model for carrier mobility has been suggested, where surface roughness, remote Coulomb and phonon scattering as well as neutral defects, ballistic transport and saturation velocity are accounted for. Then linear drain current formulation has been introduced, based upon proposed mobility, threshold voltage and inversion carrier concentration formulations. We have then reviewed the major saturation effects, as pinch off for long channel transistor and velocity saturation, injection velocity and ballistic transport for short channel devices. A compact model of saturation drain current that accounts for theses phenomena has been proposed afterward. In real devices, compact models have to account for access resistance. Hence this aspect has been treated and analytical compact models for linear and saturation regimes have been adapted. Further extraction will thus be based on the following formulation for linear drain current:

$$Id_{lin} = \frac{V_{DS}}{R_{tot}} \tag{93}$$

where the width normalized total transistor resistance R_{tot} is:

$$R_{tot} = R_0 + \frac{\sigma}{V_{GS} - V_{tLDR}} + \frac{W\left(1+\frac{L_c}{L}\right)}{\beta}\left(\frac{1}{\left(V_G - V_t - \frac{V_{DS}}{2}\right)} + \theta_1 + \theta_2\left(V_G - V_t - \frac{V_{DS}}{2}\right)\right) \tag{94}$$

The total resistance is simply the sum of contact and source-drain resistance represented by R_0 term, the LDR resistance and the channel resistance.

Saturation drain current is expressed as:

$$Id_{sat} = \frac{Id'_{sat}}{1 + G_m \cdot R_S} \tag{95}$$

where $R_s = \frac{R_0 + \frac{\sigma}{V_{GS} - Vt_{LDR}}}{2}$, Id'_{sat} is the intrinsic saturation drain current:

$$Id'_{sat} = \frac{W}{L + L_c}\mu_{eff} C_{ox}\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)V_{Dsat} \tag{96}$$

G_m is the V_G derivative of Id'_{sat}:

$$G_m = 4\frac{W}{L + L_c}\mu_{eff} C_{ox} V_{Dsat} \cdot \left(A - \theta_2\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)^2\right) \tag{97}$$

where $A = 1 + \frac{1}{L}\left(\frac{V_{Dsat} \cdot \mu_0}{v^*}\right)$ and μ_{eff} is the effective mobility as described in **(78)**, accounting for scattering mechanisms, velocity saturation and ballistic transport:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta_1\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right) + \theta_2\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)^2 + \frac{V_{Dsat}\cdot\mu_0}{Lv^*}} \quad (98)$$

and V_{Dsat} is the drain saturation voltage and is derived as V_{DS} such as $\frac{dId_{lin0}}{dV_{DS}} = 0$ where Id_{lin0} is the intrinsic linear drain current. V_{Dsat} yields:

$$V_{Dsat} = 2\frac{u - \sqrt{u.\left(1 + 2.\frac{\mu_0}{v^*.L}(V_G - V_{tsat})\right)}}{\theta_1 - \frac{2\mu_0}{L.v^*} + (V_G - V_{tsat})\theta_2} \quad (99)$$

where $u = 1 + \theta_1(V_G - V_{tsat}) + \theta_2(V_G - V_{tsat})^2$. These formulations are valid in strong inversion regime. Advantages of this formulation are that it is analytical and fast to compute. This is required to make it applicable at industrial scale and to extract parameters on a large amount of devices with very limited measured data. However, simplifications have been required in order to meet these constraints. Impacts of these simplifications have been shown to be acceptable compared to numerical calculations.

Previous formulations of drain current are reformulated here without the contribution of θ_1 and θ_2:

- For linear regime (from (94)):

$$R_{tot} = R_0 + \frac{\sigma}{V_{GS} - V_{tLDR}} + \frac{W\left(1 + \frac{L_c}{L}\right)}{\beta\left(V_G - V_t - \frac{V_{DS}}{2}\right)} \quad (100)$$

- For saturation regime (from (95), (97) and (99)):

$$Id_{sat} = \frac{Id'_{sat}}{1 + G_m \cdot R_S} \quad (101)$$

$$G_m = 4\frac{W}{L + L_c}.\frac{\mu_0}{1 + \frac{1}{L}\left(\frac{V_{Dsat}\cdot\mu_0}{v^*}\right)} C_{ox}V_{Dsat}.A \quad (102)$$

$$V_{Dsat} = L.\frac{v^*}{\mu_0}\left(\sqrt{1 + 2\frac{(V_G - V_t)\mu_0}{L.v^*}} - 1\right) \quad (103)$$

Chapter 3 :

Compact modeling: Extraction procedure and application to TCAD simulations

The aim of the compact model developed in previous chapter is to model the drain current in order to investigate device properties and their relationships with process parameters. To do so, the model will be used to fit silicon measurements and TCAD simulations in chapter 4. This step requires a method to extract model parameters. This method is detailed in paragraph §3.1. It will be applied on both TCAD simulations (in order to validate the model and the extraction method) and silicon measurements (in order to investigate devices properties). Different types of measurements are regularly performed on silicon. Modeling team performs comprehensive characterizations of transistors in order to calibrate their model. Since these measurements are time consuming, they are performed on few lots only. These characterizations include full I_D-V_G measurements on devices with different gate lengths. In the meantime, Parametric Tests (PT) are performed in line for every wafer of every lot, each wafer being measured on 17 sites. These parametric tests consist in a reduced number of measurements (few drain current measurements) in order to reduce the measurement time down to a reasonable threshold.

In order to ensure that the extraction method is robust and reliable, we use full I_D-V_G measured on silicon. Accuracy of the extraction method is check considering the fitting quality and the uncertainty about model parameters. This is the purpose of paragraph §3.2.

Monitoring process fluctuations using parameter extraction requires measurements that include all wafer and all lots. Thus full I_D-V_G characterizations from modeling team cannot be used for this purpose. Instead we will use PT. Extraction method being validated using full I_D-V_G, the influence of data sample size and range is then tested before any application on PT data. These tests are gathered in §3.3. They include tests about the influence of noise on measurements as well. Their purpose is to evaluate the uncertainty about extracted parameters depending on the considered sample size, range and noise level. It brings insights into the data amount and quality required to ensure a proper extraction.

The extraction method being set, it is then applied on TCAD simulations in §3.4. A design of experiment has been simulated in order to investigate the influence of critical process parameters on model parameters. To this end, simulation results are used for model parameters extraction. Model parameters responses to process variations are investigated and we will see if they are consistent with the physics underlying model parameters, as introduced in chapter 2.

3.1 Model parameter extraction method

In the previous chapter we have derived the linear drain current equation and we have seen that there are a limited number of model parameters that are to be extracted (R_0, σ, $\mu_0.C_{ox}$, θ_1, θ_2, V_t and V_{tLDR}, L_c). Linear drain current equation is recalled here for convenience:

$$Id_{lin} = \frac{V_{DS}}{R_{tot}} \tag{104}$$

With R_{tot} expressed as:

$$R_{tot}.W = R_0 + \frac{\sigma}{V_{GS} - V_{tLDR}} + \frac{L + L_c}{\mu_0 C_{ox}} \left(\frac{1}{\left(V_G - V_t - \frac{V_{DS}}{2}\right)} + \theta_1 + \theta_2 \left(V_G - V_t - \frac{V_{DS}}{2}\right) \right) \tag{105}$$

Model parameters will be extracted based on drain current data. This section describes the extraction method that will be used. In term of model parameter extraction, two different approaches exist: direct

extraction (involving linear regression, derivation or integration of electrical characteristics such as C_G-V_G and I_D-V_G) and nonlinear optimization algorithms where an objective function, that is the error between model and measurements, is being minimized as the model parameters converge toward there optimal value through an iterative procedure. Our approach relies first on linear least square fit in order to get a first approximation of model parameters. Then these values are used as input to a nonlinear optimization algorithm to refine their value. §3.1.1 and §3.1.2 explain the procedure to extract Vt_{lin} and the other linear parameters respectively, using least square regression. §3.1.3 and §3.1.4 details the nonlinear optimization algorithm used to finalize the extraction in linear and saturation regime respectively.

3.1.1 Threshold voltage

In literature, the threshold voltage definition is quite ambiguous and there are at least as many extraction methods as definitions of V_t available. Most common ones are Hamer's method [92], constant current or constant charge method [93], I_D-V_G linear extrapolation [94], maximum of transconductance derivative [95] that is equivalent to the maximum of capacitance derivative [96] and the Y function [97]. A comprehensive study of extraction methods has been done by Ortiz-Conde [98]. For our purpose we will choose a method that is suitable considering linear drain current equations (104)-(105) and that requires as less measurement points as possible. Every method mentioned above requires full I_D-V_G or C_G-V_G measurements in order to extract the threshold voltage (making them unsuitable for process monitoring at industrial scale) except Hamer's method. Thus, in the following section, we will adapt Hamer's method to extract the threshold voltage that suits equations (104)-(105).

Hamer's method brings the solution of a, b and c considering the following equation:

$$Z_n = a\frac{X_n - b}{Y_n - c} \tag{106}$$

where Z_n are the drain currents and $X_n = Y_n$ are the gate voltages with $n \in [1;3]$. If we consider the case where R_{sd} is constant and $\theta_2 = 0$, then I_{Dlin} (104) can be rearranged to match (106):

$$I_{Dlin} = \frac{\beta W V_{DS}}{\theta_1 + R_{sd}\beta}\frac{V_{GS} - V_t - \frac{V_{DS}}{2}}{V_{GS} - V_t - \frac{V_{DS}}{2} + \frac{1}{\theta_1 + R_{sd}\beta}} \tag{107}$$

Then $a = \frac{\beta W V_{DS}}{\theta_1 + R_{sd}\beta}$, $b = V_t + \frac{V_{DS}}{2}$, and $c = V_t + \frac{V_{DS}}{2} - \frac{1}{\theta_1 + R_{sd}\beta}$. Knowing b we can then deduce V_t. The other parameters cannot be extracted though since there are 4 unknowns (V_t, θ_1, R_{sd} and β) and only 3 equations provided by Hamer's method. If now we consider that R_{sd} depends on V_{GS} with $V_{tLDR} = V_t$ then I_{Dlin} equation (104)-(105) can be rearranged to match (106) as:

$$I_D = \frac{\beta W V_{DS}}{\theta_1 + R_0\beta}\frac{V_{GS} - V_t - \frac{V_{DS}}{2}}{V_{GS} - V_t - \frac{V_{DS}}{2} + \frac{1 + \beta\sigma}{\theta_1 + R_0\beta}} \tag{108}$$

and $a = \frac{\beta W V_{DS}}{\theta_1 + R_0\beta}$, $b = V_t + \frac{V_{DS}}{2}$, and $c = V_t + \frac{V_{DS}}{2} - \frac{1+\beta\sigma}{\theta_1 + R_0\beta}$. Then knowing b, V_t can be easily deduced. Although this adaptation of Hamer's method is not the state-of-the-art method to extract threshold voltage, this approach is perfectly consistent with our formulation of drain current if θ_2 is neglected

and Vt_{LDR} equals Vt_{lin}. This consistency is the critical point for this kind of task as discussed by McAndrew and Layman [99].

If $V_{tLDR} \neq V_t$ and $\theta_2 \neq 0$, Hamer's method cannot be used to extract accurately V_{tlin} but it can be used to have an approximation of its value. In order to get an accurate value of V_{tlin} and V_{tLDR}, a nonlinear optimization algorithm will be used as a successive step of linear least square regression.

3.1.2 Access resistance, effective channel length and mobility extraction

In this section, a review of published extraction method based on the same linear drain equation as (104)-(105), knowing Vt_{lin}, is reported. Their differences will be analyzed in order to determine the method to be used for our purpose.

In our model, the concept of access resistance is linked to parameters R_0 and σ. In contrast, $\mu_0.C_{ox}$, θ_1, θ_2, and Lc are related to channel conductivity. In order to extract these parameters we review the different methods published in the literature. Brews [100] and McAndrew [101] made a review of the main characterization methods (up to eleven of them) to extract $R_{SD}, L_{eff}, V_t, \beta$ and θ_1. Among them we find Suciu and Johnston [102], peak gm [103][104], $1/\beta$, R_{lin}-L [105][106] (that is similar to TMC [107][108]), De La Moneda [109], Peng [110], Sheu [88], Peng and Afromowitz [112], Whitfield [113] and Chang and Berg [114] methods. To that list we can add Ghibaudo's Y function [115]-[120], Taur's shift and ratio [121], Biesemans [122], Sanchez [123] and Jeppson and Karlsson's [124][125] methods. All these methods are direct extraction methods based on derivatives and linear regression of I_D-V_G characteristics. The difference between them lies in the assumptions made beforehand and the regression they use to extract parameters, but drain current equations are equivalent. Table 3-1 provides a summary in term of assumptions and extracted parameters.

Table 3-1: Model parameter extraction method summary

Method	*Assumptions*	*Extracted parameters*
$1/\beta$	$R_{SD} = 0$, μ constant with L	$\mu(V_G)$, ΔL
Chang and Berg [114]	$R_{SD} = 0$, μ constant	μ, ΔL
Suciu and Johnston [102] Jeppson and Karlsson's method [124][125] Sanchez [123] peak g_m [103][104] De La Moneda [109]	R_{SD} constant, $\mu = \frac{\mu_0}{1+\theta_1(V_G-V_t)}$ μ_0 constant with L	R_{SD}, μ_0, θ_1, ΔL
R_{lin}-L [105][106] Sheu [88]	$R_{SD} = R_0 + \frac{\sigma}{V_G - V_t}$ $\mu = \frac{\mu_0}{1+\theta_1(V_G-V_t)}$ μ constant with L, L constant with V_G.	σ, μ_0, θ_1
Hu [106]	μ constant	$R_{SD}(V_g)$, L_{eff}
TMC [107][108]	R_{SD} constant, μ constant with L	$(V_G - V_t).\mu(V_G)$, ΔL, R_{SD}
Peng [110]	R_{SD} constant, μ constant with L	$\mu(V_G)$, L_{eff}, R_{SD}
Peng and Afromowitz [112]	R_{SD} constant, $I_D.R_{SD} << V_{DS}$,	R_{SD}, ΔL

	μ constant with L	
Whitfield [113]	R_{SD} constant, μ constant with L	ΔL, R_{SD}
Y function [115]-[120]	R_{SD} is constant in the first papers. Generalization of the method canceled this assumption.	$R_{SD}(V_G)$, $\mu_0(V_G, L)$, θ_1, θ_2.
Taur's shift and ratio [121]	R_{SD} constant, L_{eff} independent of V_{GS}	R_{SD}, L_{eff}
Brut [126], Yamaguchi[127], Biesemans[122]	μ constant with L	$R_{SD}(V_G)$, $L_{eff}(V_G)$

The two first methods of Table 1-1 have inappropriate assumptions since nowadays, access resistance represents up to 70% of the total resistance for shortest devices [128]. In line 3, all the proposed methods assume a constant access resistance and extract ΔL whereas line 4 (to which our approach belongs to if it is assumed that V_{tLDR}=V_{tlin} and $\theta_2 = 0$) assumes $R_{SD} = R_0 + \frac{\sigma}{V_G - V_t}$ and L constant. However, these formulations are equivalent. Indeed, in linear regime, the width normalized total resistance formula yields:

$$R_{lin} = R_0 + \frac{\sigma}{V_{GS} - V_{tLDR}} + \frac{L + L_c}{\mu_0 C_{ox}} \left(\frac{1}{\left(V_G - V_t - \frac{V_{DS}}{2}\right)} + \theta_1 + \theta_2 \left(V_G - V_t - \frac{V_{DS}}{2}\right) \right) \quad (109)$$

In this expression, L_c is not distinguishable from ΔL. So even if their physical meaning is different, these parameters are the same considering extraction. In the particular case where $V_{tLDR} = V_{tlin}$ and θ_2 is neglected, σ cannot be distinguished from L_c either. Indeed, R_{lin} expression (109) where Lc=0 and $\theta_2 = 0$ can be rearranged as:

$$R_{lin}.W = R_0 - \sigma\theta_1 + \frac{L + \sigma\mu_0 C_{ox}}{\mu_0 C_{ox}} \left(\frac{1}{\left(V_{GS} - V_t - \frac{V_{DS}}{2}\right)} + \theta_1 \right) \quad (110)$$

In this expression $R_0 - \sigma\theta_1$ is similar to the constant access resistance term and $L_c = \sigma\mu_0 C_{ox}$. Thus, extracting R_{sd} and ΔL using method of line 3 is equivalent to extract $R_0 - \sigma\theta_1$ and $\sigma\mu_0 C_{ox}$ using methods of line 4. Thus these methods only differ from the extraction method used. This equivalence of σ and ΔL makes simultaneous extraction of $R_{SD}(V_{GS})$ and $L_{eff}(V_{GS})$, as proposed by Hu [106], Brut [126] and Yamaguchi [127] meaningless unless a clear and physical definition of the channel length is provided (involving for example critical carrier density as mentioned by Biesemans [122] or the metallurgical junction as mentioned by Lou [129]). In other words, if Vt_{LDR}=V_{tlin}, using linear regression, these parameters can't be distinguished mathematically and any workaround would yield highly correlated parameter values as demonstrated by Brut [126]. TMC [107][108], Peng and Afromowitz [112], Taur's shift and ratio [121] and Whitfield [113] methods at least assumes that R_{SD} is constant, thus their model are equivalent to ours.

More recent publications have proposed iterative procedures in order to extract both $Rsd(V_{GS})$ and $\mu(V_{GS})$ (Fleury [118][119] and Subramanian [120]) or $Rsd(V_{GS})$ and $\Delta L(V_{GS})$ (Kim [91]). However we have seen that Lc accounts for both ΔL and μ(L) roll down. Moreover, there is an equivalence between Lc and σ if V_{tlin}=V_{tLDR} and $\theta_2 = 0$. Thus extracting both R_{sd} and μ (or ΔL) depending on the gate

voltage, without alleviating the ambiguity about the access/channel region splitting, leads to results that are strongly dependent on the initial guess (that will determine the access/channel region splitting).

Thus for our extraction we will test whether L_c and σ can be distinguished (depending on the discrepancy between V_{tLDR} and V_{tlin}). If they are, equation (105) will be used for parameter extraction. If not, the equation will be simplified (by either using only σ or L_c or even removing both of these terms if access resistance is constant and Lc close to 0).

3.1.3 Linear model parameter extraction method

Let us now introduce the method of our own [105][128][131]. The extraction procedure is based on the linear drain current equation (105). This formulation does not allow any linearization for a direct extraction procedure and Hamer's method cannot be used "as is" as discussed in §3.1.1. In order to alleviate this difficulty, we first consider $L_c = 0, V_{tLDR} = V_{tlin}$ and $\theta_2 = 0$ and extract Vt_{lin} using Hamer's method. Then every other parameter is extracted at once using the following system of linear equation [131]:

$$R_{lin}(Vg, L) = \begin{pmatrix} 1 & \frac{1}{V_{gt}} & \frac{L}{V_{gt}} & L & L \cdot V_{gt} \end{pmatrix} \cdot \begin{pmatrix} R_0 \\ \sigma \\ \frac{1}{\mu_0 \cdot C_{ox}} \\ \frac{\theta_1}{\mu_0 \cdot C_{ox}} \\ \frac{\theta_2}{\mu_0 \cdot C_{ox}} \end{pmatrix} \quad (111)$$

This first step yields an initial guess for the model parameters. Then we use it as input to a nonlinear optimization method to extract the suited values of $R_0, \sigma, \theta_1, \theta_2, \mu_0 \cdot C_{ox}, V_{tLDR}$, L_c and V_{tlin}. Nonlinear optimization algorithms have been first used by McAndrew [101] who showed that this approach is more robust than typical direct extraction method. It is nowadays widely spread and used for complex compact model parameter extraction such as BSIM4 [132] or PSP [133]-[135]. Nonlinear optimizer we use is a built-in Matlab function based on trust-region-reflective and conjugate-gradient algorithm [136]-[138]. An excerpt of the code is available in Appendix A.

The strength of nonlinear optimization method compared with direct extraction methods (apart from being able to handle nonlinear problem) is that they are less sensitive to ill-conditioned problems. In our case, depending on the chosen V_{GS} and L values, column two, three and four of equation (111) matrix can be more or less correlated. For extreme cases, this can lead to singularity of the matrix and make (111) unsolvable. The other drawback of linear least square problem (111) is that it minimizes the square of the difference between measured and modeled R_{lin} and this resistance is larger for long transistors than short ones. Thus, extracted parameters will advantageously fit long transistors at the expense of the short ones (that are actually the one of interest), biasing the extraction. On the contrary, nonlinear optimization methods can either optimize the drain current model error or the resistance model error. It can even be used to optimize the normalized error between model and measurements, leading to a uniform model error across the whole range of V_G and L data.

3.1.4 Saturation model parameter extraction method

This section explains the method used to extract model parameters in saturation regime. Saturation drain current introduced in chapter 2 is recalled here:

$$Id_{sat} = \frac{Id'_{sat}}{1 + G_m \cdot R_S} \tag{112}$$

where $R_s = \frac{R_0 + \frac{\sigma}{V_{GS} - Vt_{sat}}}{2}$ and Id'_{sat} is the intrinsic saturation drain current:

$$Id'_{sat} = \frac{W}{L + L_c} \mu_{eff} C_{ox} \left(V_{GS} - V_t - \frac{V_{Dsat}}{2} \right) V_{Dsat} \tag{113}$$

G_m is the V_G derivative of Id'_{sat}:

$$G_m = 4 \frac{W}{L + L_c} \mu_{eff} C_{ox} V_{Dsat} \cdot \left(A - \theta_2 \left(V_{GS} - V_t - \frac{V_{Dsat}}{2} \right)^2 \right) \tag{114}$$

where $A = 1 + \frac{V_{DS} \cdot \mu_0}{L v^*}$ and μ_{eff} is the effective mobility, accounting for scattering mechanisms and velocity saturation:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta_1 \left(V_{GS} - V_t - \frac{V_{Dsat}}{2} \right) + \theta_2 \left(V_{GS} - V_t - \frac{V_{Dsat}}{2} \right)^2 + \frac{V_{Dsat} \cdot \mu_0}{L \cdot v^*}} \tag{115}$$

and V_{Dsat} is the drain saturation voltage and is derived as the V_{DS} value such as $\frac{dId_{lin0}}{dV_{DS}} = 0$ where Id_{lin0} is the intrinsic linear drain current. V_{Dsat} yields:

$$V_{Dsat} = 2 \frac{u - \sqrt{u \cdot \left(1 + 2 \cdot \frac{\mu_0}{v^* \cdot L} (V_G - V_{tsat}) \right)}}{\theta_1 - \frac{2\mu_0}{L \cdot v^*} + (V_G - V_{tsat}) \theta_2} \tag{116}$$

where $u = 1 + \theta_1 (V_G - V_{tsat}) + \theta_2 (V_G - V_{tsat})^2$.

Saturation drain current equation is not linearizable. Thus nonlinear optimization is used as the only step for extracting saturation parameters. Saturation parameters are $C_{ox} \cdot v^*$ and V_{tsat}. Hopefully v^* is a fairly stable parameter and its value is close to saturation velocity or injection velocity. Both of these quantities have been accurately measured and reported in literature and are very close to each other. Thus we can safely use it as first guess for the optimizer. First guess for v^* is set to 10^7cm/s for nMOS and 6.10^6 cm/s for pMOS. C_{ox} is calculated knowing the equivalent gate oxide thickness deposited during the process. For nMOS, $C_{ox} = 3.21 \cdot 10^{-6} F/cm^2$ and for pMOS $C_{ox} = 2.84 \cdot 10^{-6} F/cm^2$. V_{tlin} is chosen as the first guess for V_{tsat} extraction.

3.1.5 Summary of the extraction method

To sum up, the extraction procedure starts with linear regime measurements. R_0, σ, $\mu_0 . C_{ox}$, θ_1, θ_2 and V_{tlin} parameters are first approximated using a linear least square regression and Hamer's method.

Then these values are used as a first guess of a nonlinear optimizer to refine these model parameters and extract L_c, V_{tLDR}. Based on these extracted parameters, saturation model parameters v^* and V_{tsat} are extracted using the same nonlinear optimizer. Values from literature and V_{tlin} are taken as first guess of v^* and V_{tsat} respectively.

3.2 Extraction on full I_D-V_G curves measured on silicon

In order to assess the functionality of the extraction method as well as the validity of the model, in this paragraph we perform extraction on silicon measurements of 28 and 14 nm FD-SOI devices. Here, extraction is performed using full I_D-V_G in strong inversion regime. Extraction method efficiency is assessed by the fitting quality and the uncertainty about extracted parameters. Depending on each device type, different gate length as well as gate biases are used to measure drain current. Data samples are detailed in Table 3-2.

Extraction results are detailed for 28 nm FD-SOI nMOS devices. Figure 3-1 shows the measured and modeled I_D-V_G curves in linear and saturation regime for every gate lengths. Extracted parameters are gathered in Figure 3-2 and Table 3-3.

Gate length [µm]			
28 FD-SOI (Silicon)		14 FD-SOI (Silicon)	
nMOS	pMOS	nMOS	pMOS
0.024	0.024	0.022	0.022
0.0276	0.0276	0.024	0.024
0.0312	0.0312	0.026	0.026
0.078	0.078	0.028	0.028
0.105	0.105	0.03	0.03
0.447	0.447	0.032	0.032
0.897	0.897	0.034	0.034
8.997	8.997	0.038	0.038
		0.042	0.042
		0.064	0.064
		0.104	0.104
		0.154	0.154
		0.164	0.164
		0.504	0.504
		1.004	1.004
		2.004	2.004
		3.004	3.004

(a)

	Gate voltage [V]			
	28 FD-SOI (Silicon)		14 FD-SOI (Silicon)	
	nMOS	pMOS	nMOS	pMOS
Min [V]	0.55	0.7	0.614	0.614
Step [mV]	27.5	20	26	26
Max [V]	1.1	1.1	0.9	0.9

(b)

Table 3-2: Gate length and biases used to measure I_{Dlin} and I_{Dsat} depending on the device type considered.

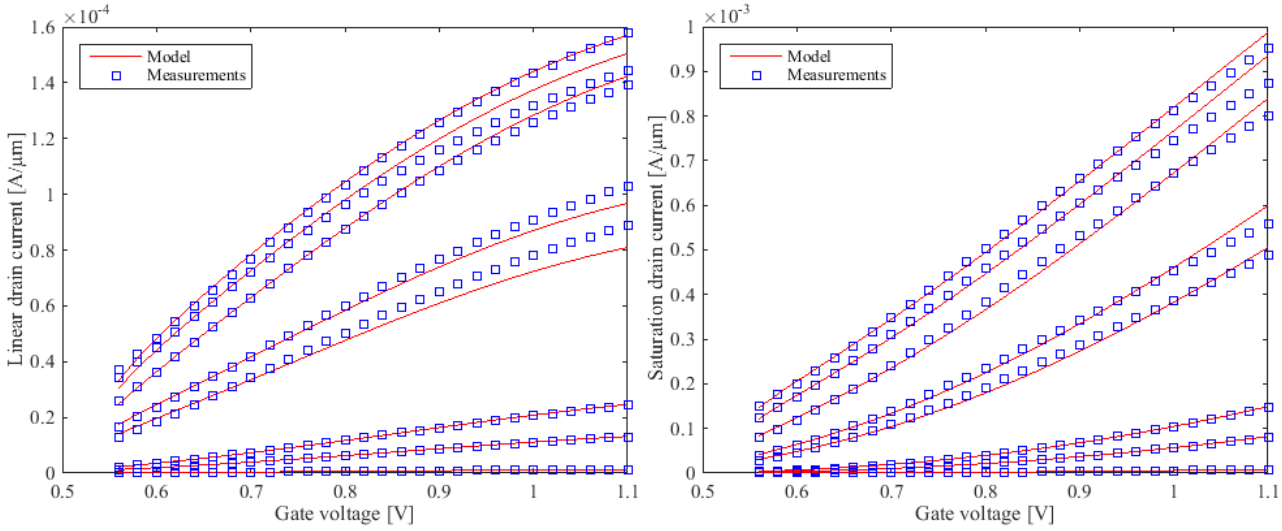


Figure 3-1: Modeled and measured 28 nm FD-SOI nMOS I_{Dlin} (a) and I_{Dsat} (b) against gate voltage for different gate length.

Parameters	Values
R_0	121 [Ω. µm]
σ	23.6 [Ω. µm. V]
$\mu_0 . C_{ox}$	$2.41 \cdot 10^{-4}$ $\left[\frac{F}{V.s}\right]$
θ_1	-1.40 [V^{-1}]
θ_2	1.23 [V^{-2}]
Vt_{LDR}	0.519 V
L_c	12.6 [nm]
$v^* . C_{ox}$	$4.55 \cdot 10^{-3}$ [F/µm/s]

Table 3-3: Extracted model parameters for 28 nm FD-SOI nMOS transistors measured on silicon.

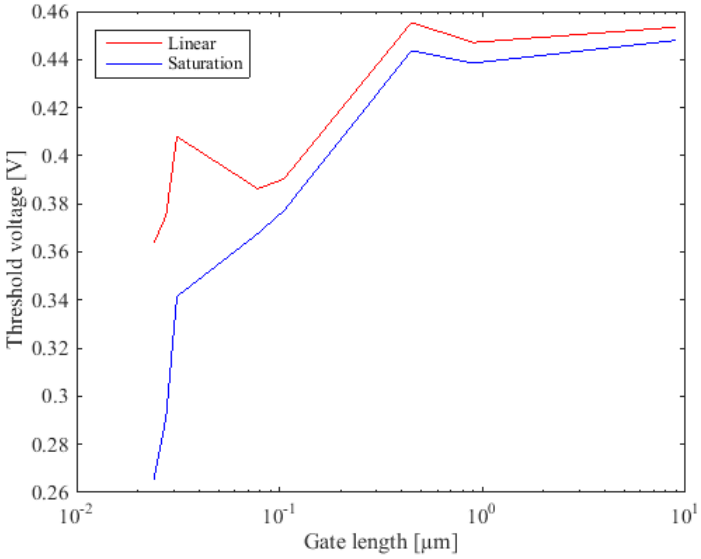
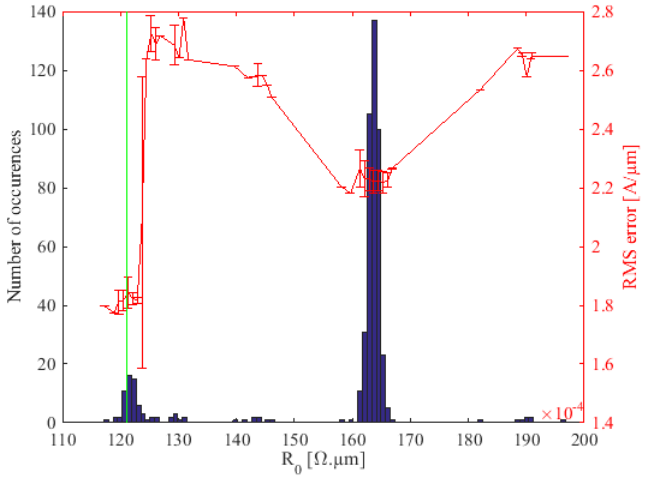
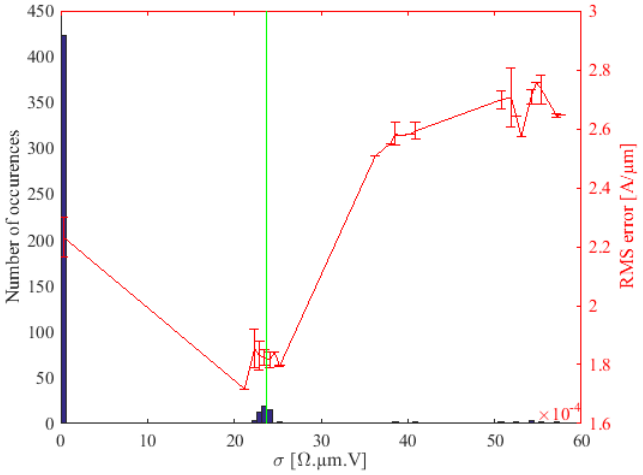


Figure 3-2: Extracted V_{tlin} and V_{tsat} against gate length for 28 nm FD-SOI nMOS devices measured on silicon.

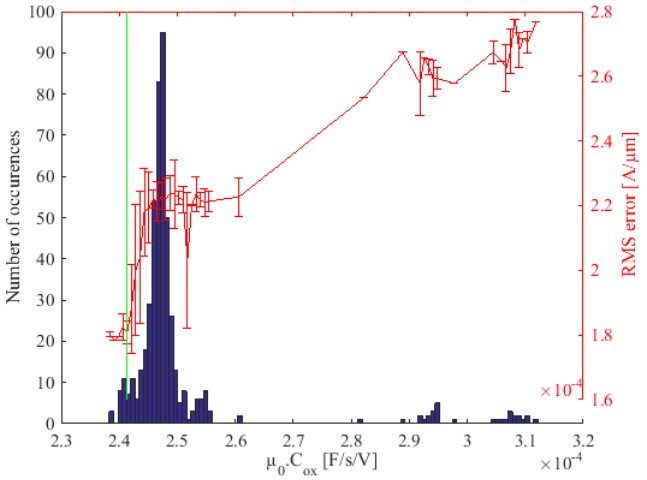
The quality of fit is appropriate. In order to estimate the uncertainty about extracted parameters and the robustness of extracted parameters, a cross validation method is applied. This method consists in withdrawing few measurements from the data sample and performing the extraction procedure again. The higher the discrepancy between the two results, the more uncertain the extraction results are. The last step can be repeated many times, each time withdrawing a different subset of data, in order to estimate this uncertainty (see chapter 5 for more information about cross validation methods). In the current case, full data sample consist in 224 measurements in linear and saturation regime. The extraction procedure is repeated 500 times, each time withdrawing a different data subset of 20 measurements. Data subsets are chosen randomly among the measurements. Histograms of extracted parameters distributions are shown in Figure 3-3.


Number of occurences
RMS error [A/µm]
R0 [Ω.µm]
×10-4

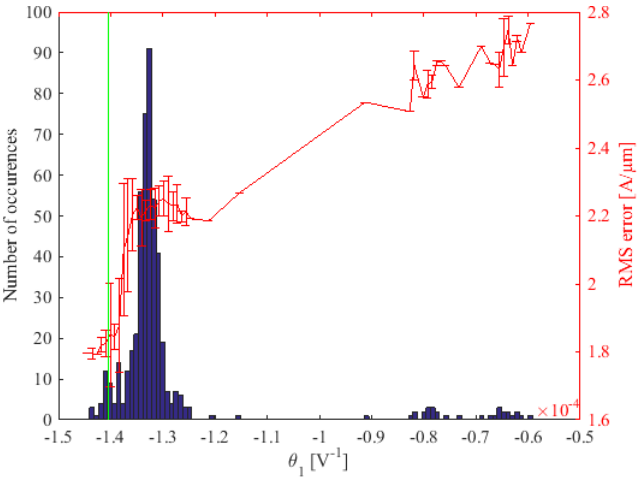
(a)


Number of occurences
RMS error [A/µm]
σ [Ω.µm.V]
×10-4

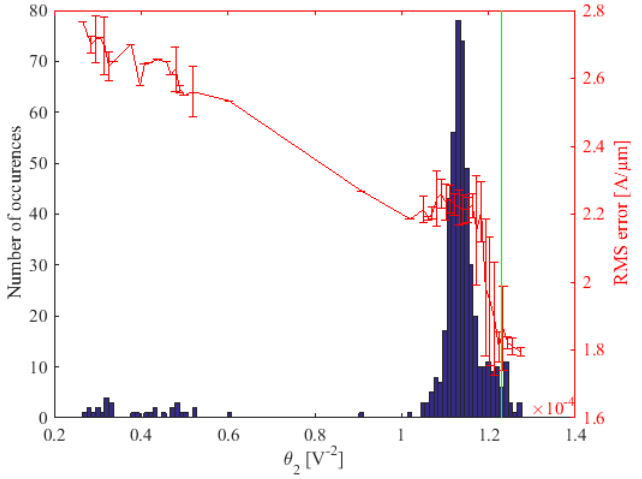
(b)


Number of occurences
RMS error [A/µm]
µ0.Cox [F/s/V]
×10-4
×10-4

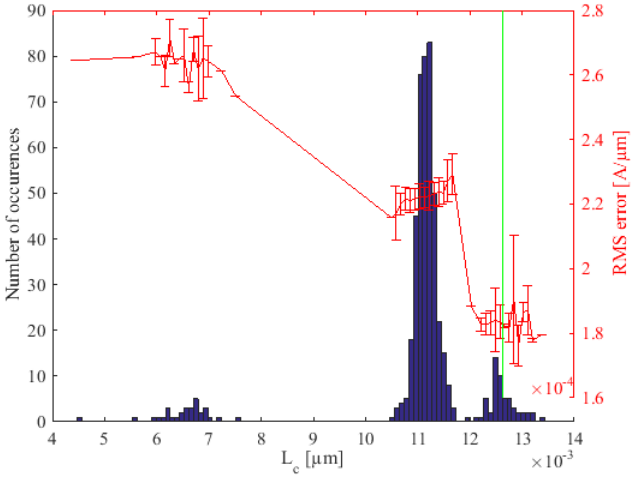
(c)


Number of occurences
RMS error [A/µm]
θ1 [V-1]
×10-4

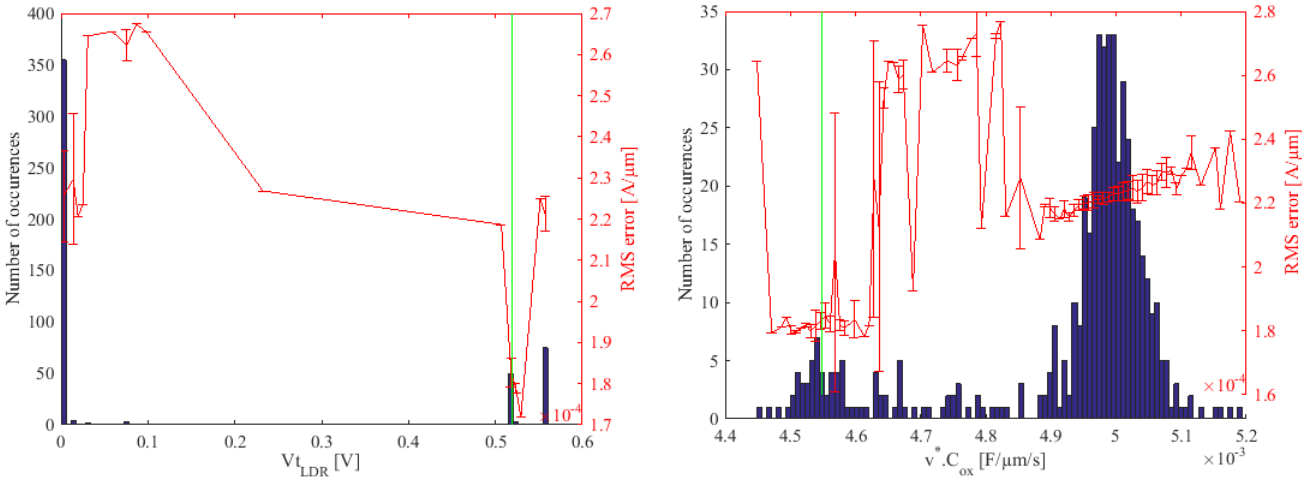
(d)


Number of occurences
RMS error [A/µm]
θ2 [V-2]
×10-4

(e)


Number of occurences
RMS error [A/µm]
Lc [µm]
×10-4
×10-3

(f)



(g) (h)

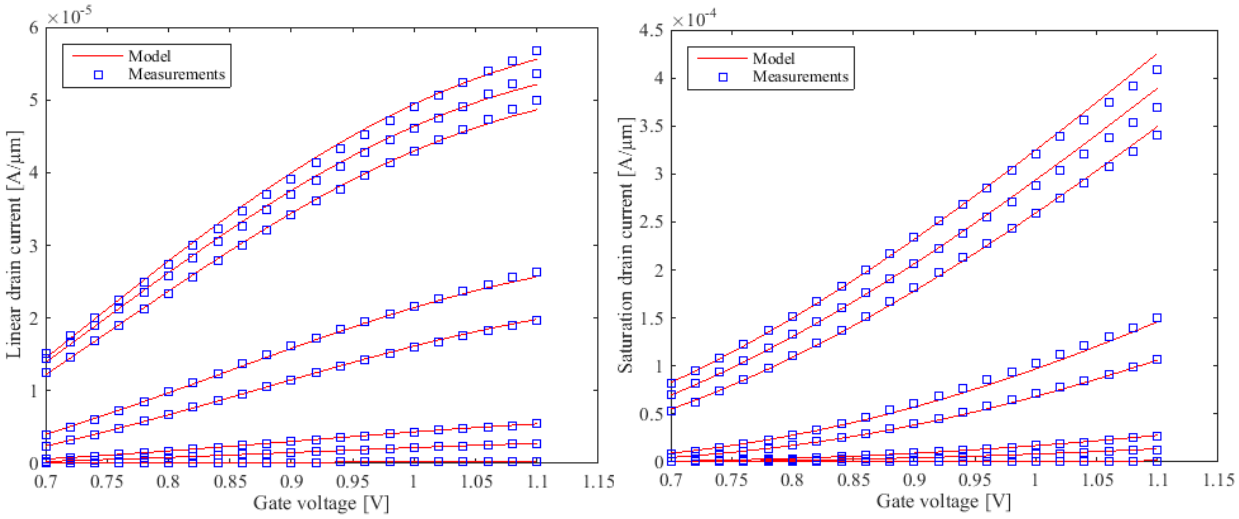
Figure 3-3: Histograms of extracted parameter distribution using cross validation method.

Along with histograms, the root mean square model error is plotted in Figure 3-3. This error is calculated using equation (117) where $Idlin_{synth}$ is the synthetized drain current and $Idlin_{model}$ is the modeled one.

$$RMS = \sqrt{\sum_i \sum_j \left[Idlin_{synth}(Vg_i, L_j) - Idlin_{model}(Vg_i, L_j)\right]^2} \quad (117)$$

This error is calculated over the full data sample (including withdrawn data subset for extraction). It evaluates the predictability of the model depending on the parameter value. Error bars represents the standard deviation of the model error. In each plot, the green line represents the value extracted using the whole data set. Studying carefully the different plots we see that results are gathered around 2 distinct solutions. The most observed solution yields $\sigma = Vt_{LDR} = 0$ and $R_0 = 163\ \Omega.\mu m$ whereas the most accurate solution (with the lowest model error) is the one found using the whole dataset for extraction and has non zeros σ and Vt_{LDR} with a lower R_0. Thus we see that in the first case the V_G dependent access resistance has been substituted by a constant access resistance. However this results is less accurate than the second where access resistance depends on the gate voltage. This V_G dependent access resistance seems touchy to extract properly. This is partly due to the fact that its influence on model error is limited since it only drops it from 220 µA/µm down to 180 µA/µm. Moreover the error is located on the short channel devices drain current at low V_G. Thus removing one measurement point from these devices can easily bias the extraction and steer the nonlinear solver to converge toward the local minimum where access resistance does not depend on gate voltage. This extraction issue will be investigated in §3.3.

As a counterexample, the case of 28 nm FD-SOI pMOS devices is studied and results shows that the access resistance does not depends on V_G and the results are robust. Measurements against model are shown in Figure 3-4 for linear and saturation regime. A good fit is obtained. Model parameters are gathered in Figure 3-5 and Table 3-4.



(a) (b)

Figure 3-4: Modeled and measured 28 nm FD-SOI pMOS I_{Dlin} (a) and I_{Dsat} (b) against gate voltage for different gate lengths.

Parameters	Values
R_0	$374\ [\Omega.\mu m]$
σ	$\sim 0\ [\Omega.\mu m.V]$
$\mu_0.C_{ox}$	$7.51 \cdot 10^{-5}\ \left[\frac{F}{V.s}\right]$
θ_1	$-1.53\ [V^{-1}]$
θ_2	$1.79[V^{-2}]$
Vt_{LDR}	~ 0 [V]
L_c	3.9 [nm]
$v^*.C_{ox}$	$4.4 \cdot 10^{-3}$

Table 3-4: Extracted model parameters for 28 nm FD-SOI pMOS transistors measured on silicon.

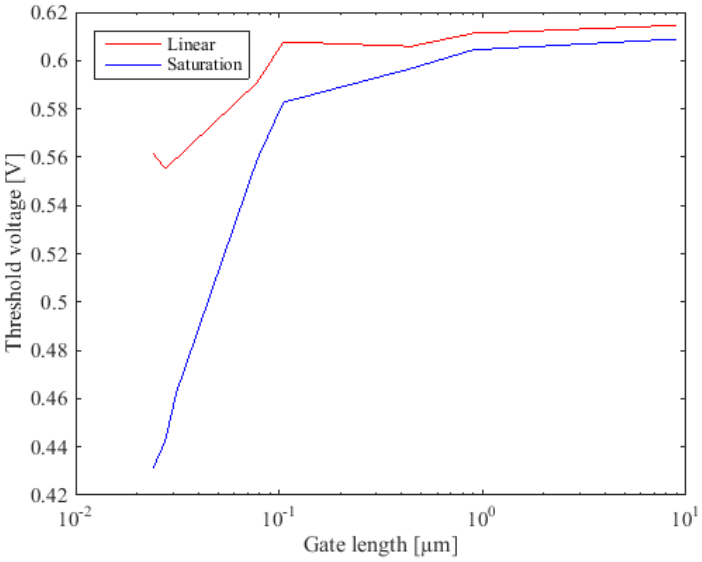


Figure 3-5: Extracted V_{tlin} and V_{tsat} against gate length for 28 nm FD-SOI pMOS devices measured on silicon.

Cross validation test results are shown in Figure 3-6. Extracted parameters are all regrouped around the solution found using the whole dataset. Model error also shows that the best solution is close to the one extracted using the whole dataset.

The reason why nMOS has an access resistance that depends on V_G contrary to the pMOS is that pMOS is overlapped and nMOS underlapped in the considered devices. This point is confirmed by TCAD simulation calibrated on 28 nm FD-SOI devices (see Figure 3-24 in §3.4). The V_G dependent access resistance region lies at both channel ends. In overlapped devices, these ends are highly doped, thus the corresponding resistance is low. This is the case of pMOS. The contrary occurs for nMOS.

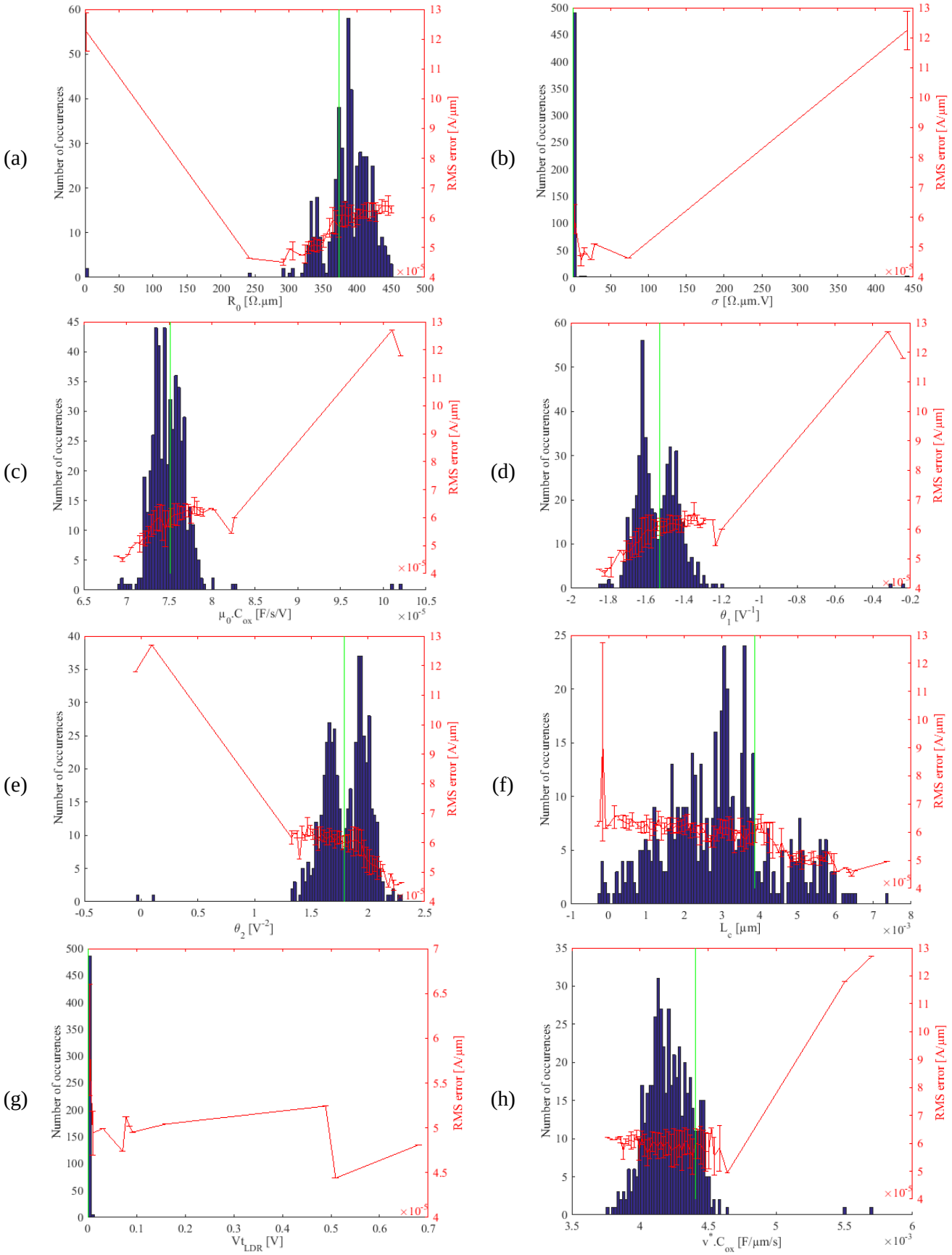


Figure 3-6: Histograms of extracted parameter distribution using cross validation method.

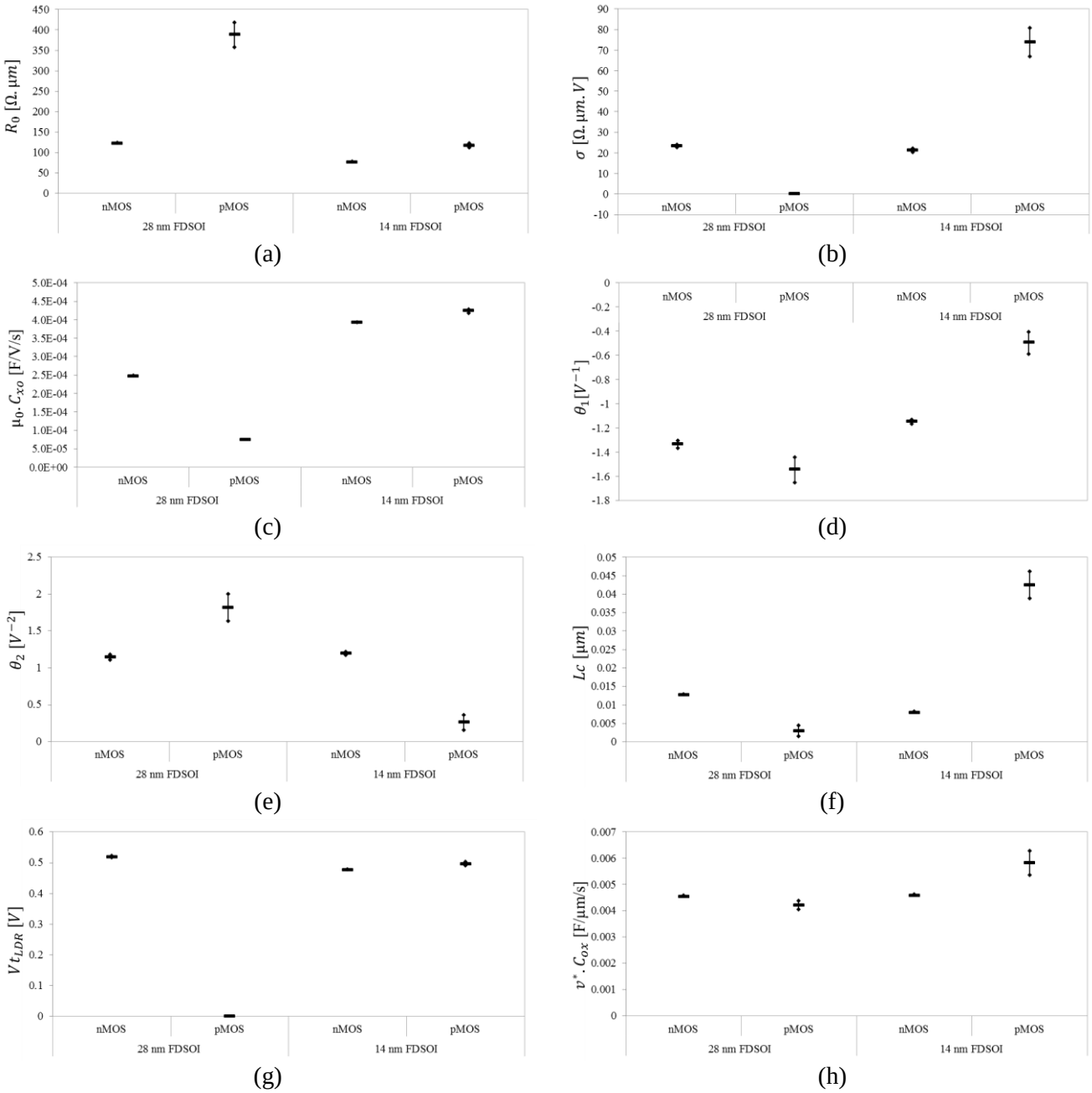


Figure 3-7: Extracted model parameters using measured full I_D-V_G for nMOS and pMOS, 28 and 14 nm FD-SOI technologies.

A summary of model parameters extracted on 28 and 14 nm FD-SOI devices using full I_D-V_G measurements is shown in Figure 3-7. In this figure, the error bars represent the parameter dispersion calculated using cross validation test. These results emphasize the fact that model parameters depend on the technology used. An enhanced mobility with lower θ parameters has been found for 14 nm FD-SOI devices. This can be due to the in-situ doped epitaxial raised source drain that induces fewer defects compared with sputtering implanted dopant used for 28 nm FD-SOI devices. It can also be due to SiGe mobility booster for the case of pMOS devices. R_0 is higher for pMOS than nMOS since hole's mobility is lower than electron's mobility and source drain are less doped for pMOS devices because of silicon solubility limit. For 14 nm FD-SOI technology, physical gate length is estimated using the drawn gate length corrected by systematic effect like OPC enabling the most accurate estimation of the physical gate length. Moreover, for 14 nm FD-SOI technology, in situ doped source drain technology prevents neutral defects formation. This is why Lc is globally close to 0. An

exception is to be noticed for 14 nm FD-SOI pMOS technology where Lc is significant. This is due to the presence of SiGe source drain boosters that induces stress in the channel. This stress is not uniform in the channel due to stress relaxation mechanisms and the average stress over the channel depends on the channel length [139][140]. Finally, as it has been discussed previously, pMOS 28 nm FD-SOI devices have no V_G dependent access resistance since they are overlapped.

To conclude we have extracted model parameters on different technologies. The model has shown its accuracy. Model parameters are dependent on the technology considered. In some cases, model parameters can be neglected. In addition, the extraction has also shown to be robust against cross validation test for the case of 28 nm FD-SOI pMOS device. However nMOS device extraction is quite unstable and two distinct solutions are found. Cross validation is required in order to discriminate the most accurate solution.

3.3 Test for extraction procedure robustness depending on data sampling

In order to monitor model parameters sensitivity to process variations, the extraction procedure should be applied on every wafer of every lot that includes process variations, using PT (spotted data) instead of full I_D-V_G curves in order to reduce measurement time. Before doing so, we must verify that the extraction procedure is still adapted using a limited amount of measurements. In this section we test the ability of the code to extract the proper model parameters values based on synthetic data depending on the data sampling. Synthetic data are artificial data created using the drain current equation and arbitrary model parameter values. This procedure is trivial but required in order to verify that:

- The method is properly implemented (no bug)
- Parameters are extractible (no redundant parameters and data sample size and range are large enough)
- The nonlinear algorithm converges properly. That is to say, it enables verifying if the termination tolerance is small enough to enable an accurate extraction of model parameters. There are two termination criterions: the minimum change in the value of the objective function during a step and the minimum size of a step in the model parameters space.

One of the major constraints of the work is that data sample size is small since PT only includes few points per curves. Data sampling is defined in §3.3.1. Thus the ability of the method to properly extract model parameters with the data sample size and range available in PT is tested in §3.3.2. Then the extraction procedure robustness is tested against artificial noise following the work done by McAndrew [101] in §3.3.3.

3.3.1 Definition of data sampling

Table 3-5 gathers the device gate lengths for which data are available depending on the technology considered. TCAD simulations have been designed such that data sample is comparable to available silicon data sample. For each of these gate lengths, drain currents has been measured in linear and saturation regimes at different gate voltages. These gate voltages are summarized in Table 3-6. The smallest data sample is available for 28 nm FD-SOI technologies where linear and saturation drain currents are measured on 6 devices with different gate lengths, each devices being measured at 3 different gate voltages. It will be refered to as “silicon data sample”.

Technology	28 nm FD-SOI	14nm FD-SOI	TCAD
Available gate lengths [µm]	0.028	0.02	0.030
	0.030	0.024	0.034
	0.034	0.03	0.038
	0.12	0.06	0.090
	0.3	0.1	0.1
	1	0.3	0.12
		1	0.15
			0.3
			1

Table 3-5: Device gate lengths for which data are available (for nMOS and pMOS).

Technology	28 nm FD-SOI		14nm FD-SOI		TCAD	
Drain bias	V_{Dlin}	V_{Dsat}	V_{Dlin}	V_{Dsat}	V_{Dlin}	V_{Dsat}
Gate voltages for which data are available [V]	0.7	0.7	Vt_{lin}+0.3	0.4	0.7	0.7
	1	1	Vt_{lin}+0.5	0.8	1	1
	1.1	1.1	0.8		1.1	1.1
			Vt_{lin}+0.7			

Table 3-6: Absolute device gate voltages for which data are available (for nMOS and pMOS).

3.3.2 Influence of data sampling

This section focuses on the influence of data sampling. We demonstrate the requirements about the model and data sample to be used in order to ensure an accurate extraction. It will be demonstrated that model extraction using synthetic data with a large data sample works. However extraction is not robust against sample range variations. We will show how to discriminate redundant parameters that can be fixed in order to improve significantly the extraction robustness without compromising the model accuracy.

3.3.2.1 Influence of data sample size and range

In order to illustrate the influence of data sample size and range, we first focus on 14 nm FD-SOI nMOS case. Results will be generalized to other technologies afterward. Synthetic data are generated using parameters close to those found for 14 nm FD-SOI nMOS devices. Values are regrouped in Table 3-7 and Figure 3-8. Parameter extraction is tested using a large data sample, similar to the one provided by full I_D-V_G measurements, harnessed in §3.2.

Synthetic data generation and model parameter extraction have been done using I_{Dlin} calculated for 20 gate lengths ranging from 22 nm up to 3 µm and 20 gate biases for each gate length ($V_G \in [0.6\ ,\ 1]$ V). First guess for nonlinear extraction is provided by linear extraction as described in §3.1 for parameters R_0, σ, $\mu_0.C_{ox}$, θ_1, θ_2 and Vt_{lin}. First guess for Vt_{LDR} and Lc are arbitrarily set to 1 mV and 1 nm. Figure 3-9 shows synthetic data and model against gate voltage. A perfect match is obtained and errors on model parameters are small. Thus in this case, data sample size is sufficient for extraction, model parameters are not redundant and the nonlinear algorithm converges properly.

Parameters	Values
R_0	76.1 [Ω. µm]
σ	21.4 [Ω. µm. V]
$\mu_0 . C_{ox}$	$3.92 \cdot 10^{-4}$ $\left[\frac{F}{V.s}\right]$
θ_1	-1.15 [V^{-1}]
θ_2	1.20 [V^{-2}]
Vt_{LDR}	0.476 [V]
L_c	7.9 [nm]
$v^* . C_{ox}$	$4.58 \cdot 10^{-4}$
V_{tlin}	$0.4 - \frac{0.1}{L^2}$
V_{tsat}	$V_{tlin} - \frac{2.5 \cdot 10^{-3}}{L}$

Table 3-7: Model parameters values for synthetic data generation. L is the gate length in nm.

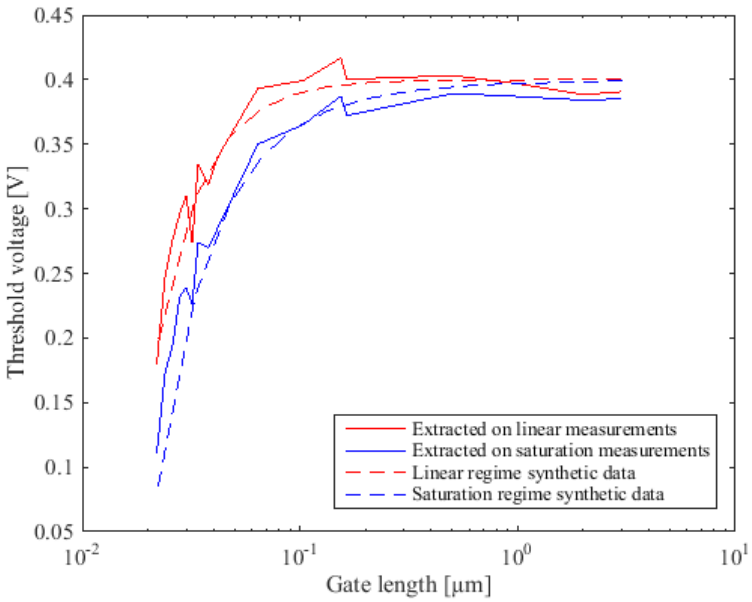


Figure 3-8: Vt_{lin} and Vt_{sat} as extracted on measurements along with those used for synthetic data generation.

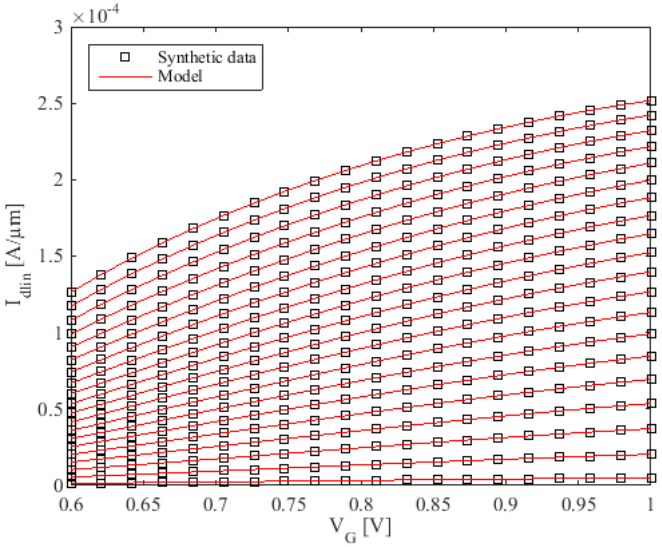


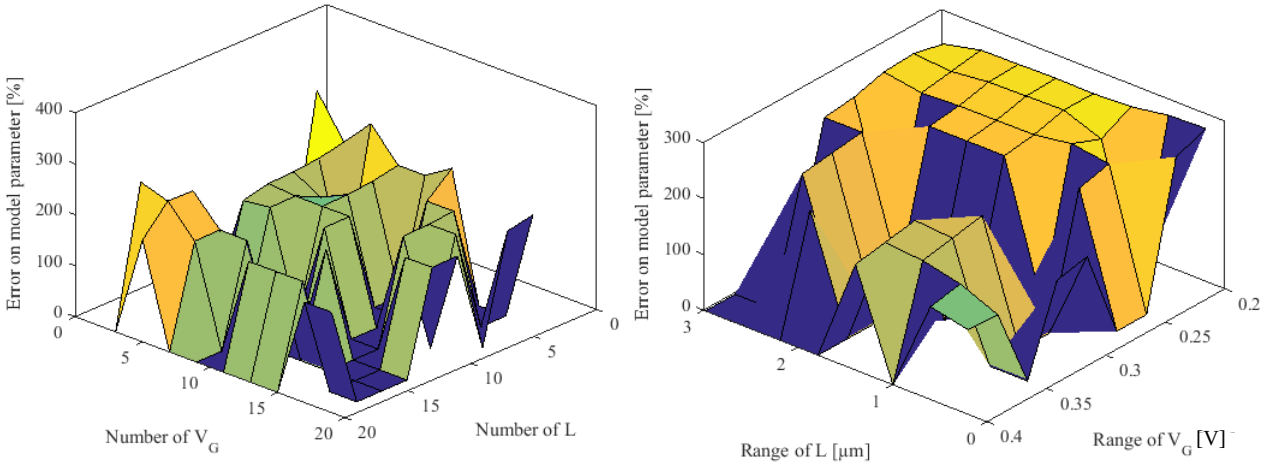
Figure 3-9: Id_{lin} modeled and synthetized against gate length, using Table 3-7 model parameters.

Model parameters	Relative error from extraction [%]
R_0	$1.98 \cdot 10^{-12}$
σ	$8.49 \cdot 10^{-12}$
$\mu_0 . C_{ox}$	$-9.69 \cdot 10^{-14}$
θ_1	$9.68 \cdot 10^{-14}$
θ_2	$1.85 \cdot 10^{-14}$
V_{tLDR}	$-6.65 \cdot 10^{-13}$
L_c	$-2.09 \cdot 10^{-11}$
V_{tlin}	$-2.8 \cdot 10^{-14}$

Table 3-8: Relative error made on model parameters from extraction.

In order to investigate the influence of the sample size and which one is required to extract properly model parameters, a test has been performed. It consists in running extraction using different sample size with a fixed range (the range used above). The size goes from the 800 data points (corresponding to 20 gate lengths and 20 gate biases in linear and saturation regimes) down to 24 data points (4 gate lengths measures at 3 different gate biases in both regimes). Figure 3-10 (a) shows the root mean square (RMS) error made on extracted model parameters against L and V_G ranges. RMS error on model parameters is calculated following **(118)** where Para is the vector of model parameters. $synth$ and $model$ subscripts stand for model parameters used as input to synthetized data and extracted model parameters respectively.

$$RMS = \sqrt{\sum_j \left[\frac{Para_{synth_j} - Para_{model_j}}{Para_{synth_j}}\right]^2} \cdot 100 \quad (118)$$



(a) (b)

Figure 3-10: Error on extracted model parameters depending on the data sampling size (a) and range (b).

In order to test whether silicon data sample range is large enough to ensure a proper extraction, extraction is performed with various sample ranges. For each sample range, the sample size is constant (20 V_G and 20 L). Results in Figure 3-10 show that the extraction is unstable and depending on the data sample considered two solutions can be found. This problem is similar to the one revealed in §3.2 when performing cross validation tests on 28 nm nMOS device extractions.

The problem mainly stems from Vt_{LDR} and Lc parameters. In fact, if Vt_{LDR} is in the range of Vt_{lin}, then it can be difficult to distinguish one from another, leading to noisy extractions. Moreover, in this case, Lc becomes hardly distinguishable from sigma as well, making them redundant (see §3.1.2 and equation (110)). Thus, in this condition, extracting σ, Vt_{LDR} and L_c leads to unstable results. To demonstrate that numerically, Figure 3-11 shows the extraction accuracy depending on the value of V_{tLDR} chosen for synthetic data generation. In this case Vt_{lin} has been set to 0.35V for all gate lengths. Figures shows that, considering this synthetic dataset, when V_{tLDR} reaches V_{tlin}, the error rises and σ, Lc and Vt_{LDR} cannot be extracted anymore if $Vt_{LDR} \geq Vt_{lin}$.

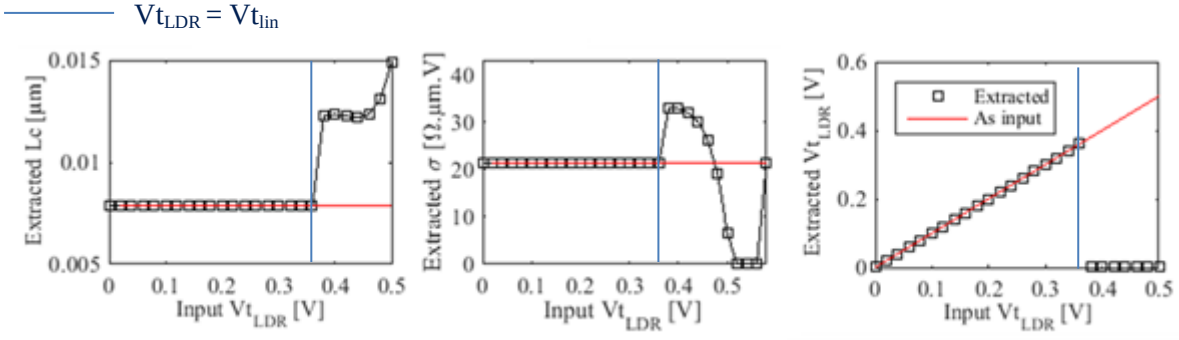


Figure 3-11: Extracted model parameters using synthetic data, depending on Vt_{LDR}.

This case study does not prove that extraction will systematically fail if Vt_{LDR} is greater or equal to Vt_{lin} but in some case, this can pose problems and cause extraction to fail. In these cases, the problem that is faced is that some parameters are partially redundant. The best way to cope with this issue is to

remove or fix the redundant parameters. However, simplifying the model leads to make approximations. In order to minimize the error created, the importance of each parameter in the model is evaluated by running test on synthetic data. The test consists in reducing the model complexity by removing each parameter one by one and calculating each time the discrepancy between the full and the reduced model. Based on model parameters obtained applying extraction on full I_D-V_G measurements on every technology, the root mean square error between full and reduced models is shown in Figure 3-12:

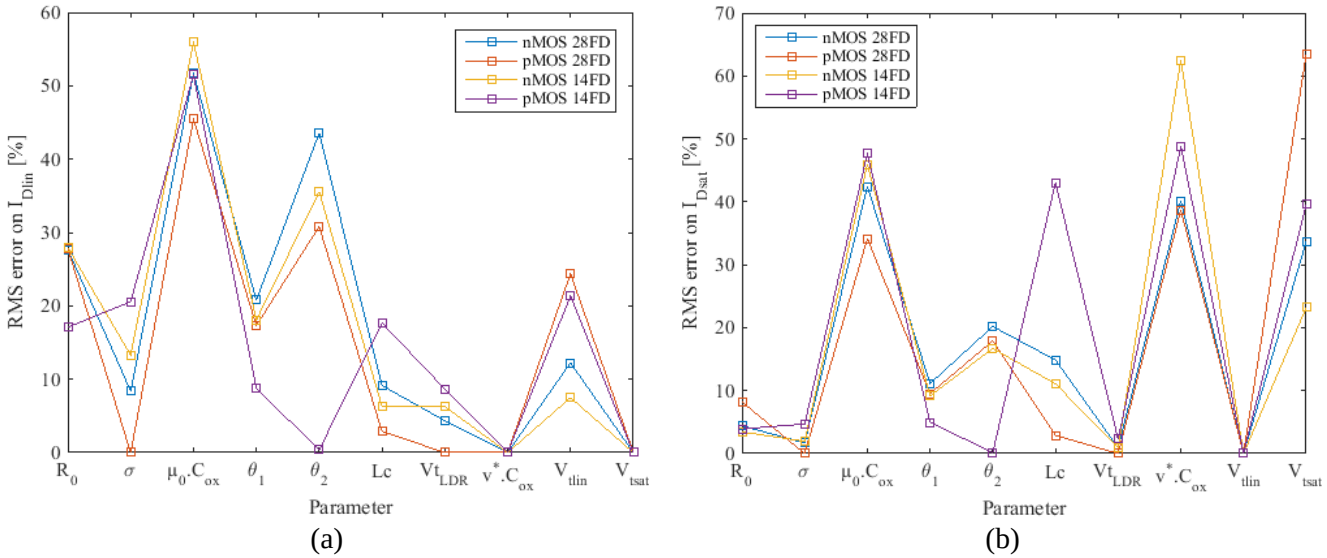


Figure 3-12: RMS error between reduced and full model, removing one parameter at a time.

Depending on the technology, model parameters are more or less important. However we see that σ, θ_1, Vt_{LDR} and Lc are the least important one considering both linear and saturation regime. Thus fixing some of these parameters will improve the robustness of the extraction procedure while inducing a minimum bias in the result. Considering the silicon data sample size and range, extractions have been performed using 5 different cases. In the first case, all parameters are considered. This is the model as described in chapter 2. In the second case, Lc has been set to 0 and is not extracted. In the third case, we assume that access resistance is constant and thus σ is set to 0. In the fourth case, access resistance is considered inversely proportional to $V_G - V_t - \frac{V_{DS}}{2}$ as suggested by Hu [106], thus Vt_{LDR} is replaced by $V_t + \frac{V_{DS}}{2}$ and is not extracted. Lc is set to 0 as well. In the last case access resistance is considered constant and Lc=0.

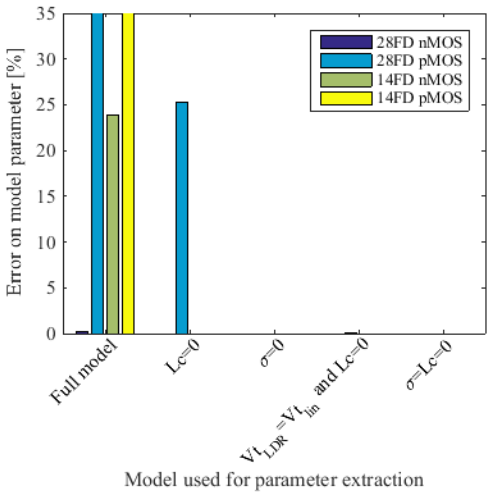
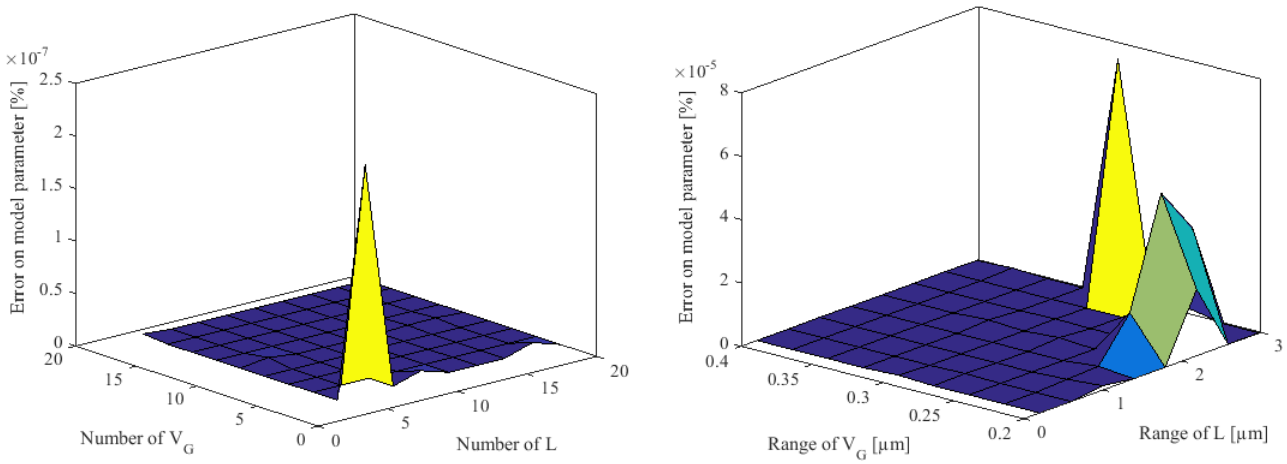


Figure 3-13: Error on extracted model parameter, using synthetic data, depending on the model used

Figure 3-13 shows the error made on extracted model parameters considering the 5 cases. Each case has been treated for each technology, using model parameters found after extraction on full I_D-V_G curves for synthetic data generation. We see that the extraction perform poorly only if every model parameters are taken into account. In addition, 28 nm FD-SOI pMOS technology is also badly extracted if only Lc is set to 0. 28 nm nMOS FD-SOI model parameters are well extracted no matter the technology considered. Thus, depending on the technology considered it may be mandatory to remove one or two parameters in order to ensure a proper extraction.

As an example, we show in Figure 3-14 the impact of sample range and size on 14 nm FD-SOI nMOS model parameters extraction, as it has been shown in Figure 3-10. However this time, Lc has been set to 0. In addition, the effect of the sample rang has been tested with only 3 gate lengths measured at 3 V_G. We see that now the error is in range of numerical noise. Thus withdrawing only one parameter can fix the problem.



(a) (b)

Figure 3-14: Error on extracted model parameters depending on the data sampling size (a) and range (b).

However, since the extraction stability depends on the technology considered, we suggeste the following procedure to ensure the extraction robustness. Every time an extraction is performed, this test with synthetic data should be run, based on extracted model parameters. A good result assesses the stability and reliability of the results.

3.3.3 Robustness against artificial noise

Conditions about the minimum data sample range and size required for a proper extraction have been set in previous paragraph. Now, considering a proper data sample for extraction, we investigate here the effect of artificial noise. To illustrate this test, we use synthetic data generated thanks to the compact model with Lc=0 and $Vt_{LDR} = V_t + \frac{V_{DS}}{2}$ and with model parameters extracted on 14 nm FD-SOI nMOS technology. These parameters are regrouped in Table 3-7 and Figure 3-8. A generalization of the method to other technologies is done afterward. In order to model noise, the drain current values are modulated by a normally distributed random amount (the noise) with dispersion (3σ) corresponding to the noise level. Model parameters are then extracted and error between the modeled and the synthetized current is calculated as well as the error between extracted model parameters and the one used as input for synthetic data generation. The extraction is performed using silicon data sample.

In practical situation, noise can arise from different sources. One source is directly linked to the measurement setup. PT uses short time measurements (few milliseconds). It means that the measure is

averaged over this period of time. Thus every high frequency noises (F > 1 MHz) are deleted. Low frequency noise (LFN) [141] only remains. In our setup and considering our technology, measurement noise does not exceed 1% of the measure. Another source that can be assimilated to "noise" is the local variability. Indeed, in our extraction, we assume that every model parameters are common for all transistors of the same die (except threshold voltage). This is only true if the local variability is neglected. In this paragraph we only focus on measurement noise. Impact of local variability effect will be treated in chapter 4 and 5.

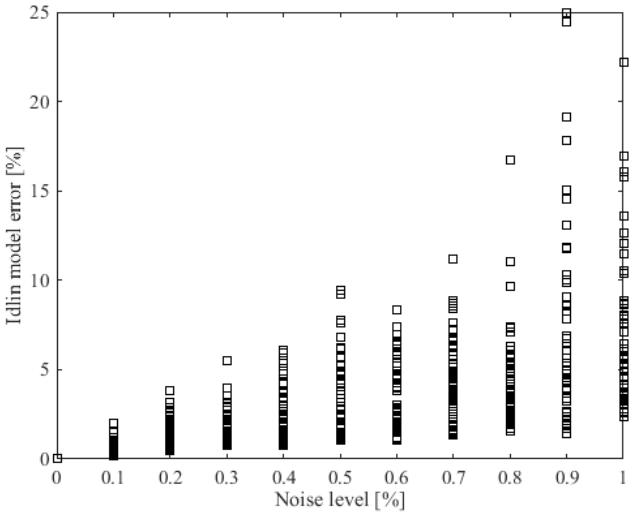


Figure 3-15: (a) RMS error against artificial noise induced in the synthetized data.

Figure 3-15 (a) shows I_{Dlin} RMS error against the noise level for each extraction. We see that the error rarely exceeds 20% with noise level up to 1%. Figure 3-15 (b) shows the worst fit obtained with 1% noise on synthetic data. This error is large and can lead to build strongly biased compact model. This is due to the limited size and range of the sample (the silicon data sample). The uncertainty about model parameter extraction mainly arises from the mobility reduction factors θ. Indeed, in linear regime (where velocity saturation and ballistic transport is neglected), the mobility compact model used is expressed as a second order expansion of $V_G - V_{tlin} - \frac{V_D}{2}$. The inverse of the mobility is thus fitted as a second order polynomial expression following:

$$\frac{1}{\mu_{eff}} = \frac{1}{\mu_0}\left(1 + \theta_1\left(V_G - V_t - \frac{V_D}{2}\right) + \theta_2\left(V_G - V_t - \frac{V_D}{2}\right)^2\right) \tag{119}$$

In order to illustrate the extraction robustness of such a compact model, we simulate the mobility against gate voltage using TCAD tool on nMOS FD-SOI device. Figure 3-16 shows the inverse of the electron mobility half way between source and drain, against the gate overdrive. TCAD simulated mobility has been averaged across the SOI layer thickness, weighted by the inversion carrier density in order to get the effective mobility. Equation used to calculate this mobility is shown below:

$$\mu_{eff}(at\ y = L/2) = \frac{\int_0^{T_{si}} \mu \cdot Q_i \cdot dx}{\int_0^{T_{si}} Q_i\, dx} \tag{120}$$

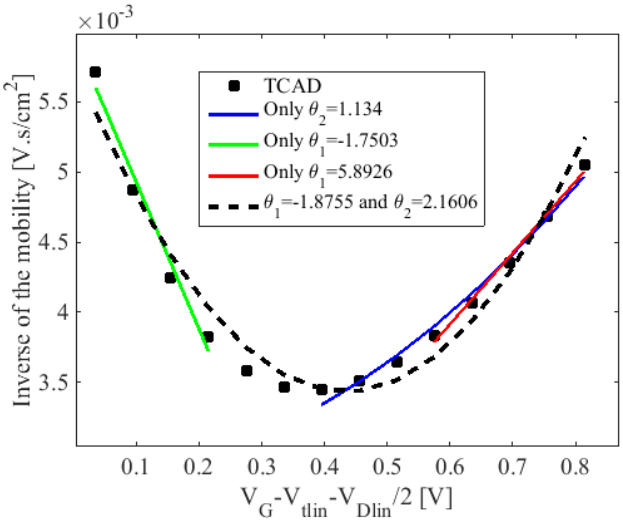


Figure 3-16: Inverse of the mobility half way between source and drain, TCAD simulated and modeled using first and second order model.

Considering the whole gate bias range of strong inversion regime, we see that the mobility has a square dependence with respect to $V_G - V_t - \frac{V_{Dlin}}{2}$. This conclusion is obvious considering the whole range of gate bias in strong inversion. However extraction will only benefit from a reduced gate bias range (0.4V) and a reduce sample size. Thus depending on the position of this range, Figure 3-16 shows that effective mobility model can be simplified using only either θ_1 or θ_2. Moreover there are 3 model parameters to be extracted and 3 drain currents measured. The problem is square but extraction results can be very noisy. Thus, considering a second order model involves too many parameters. It shall be reduced to 2 model parameters, removing either θ_1 or θ_2.

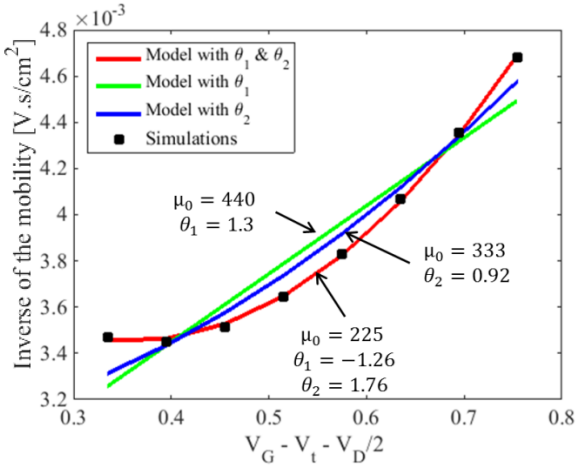


Figure 3-17: Effective mobility simulated and modeled using first and second order approximation.

Figure 3-17 shows first and second order effective mobility model along with the simulated one against gate overdrive. The range of gate overdrive is restricted to the extraction range. The best fit against the mobility curve is used to decide which parameter to remove. Model with only θ_2 is slightly better considering the fitting quality but very close to the model with only θ_1. Values of extracted μ_0, θ_1 and θ_2 are regrouped in Table 3-9 depending on the extraction range.

Parameters	Extraction using the whole Vg range	Extraction with 0.7<Vg<1.1 V using θ_1 and θ_2	Extraction with 0.7<Vg<1.1 V using θ_1	Extraction with 0.7<Vg<1.1 V using θ_2
$\mu_0[cm^2/V/s]$	173	225	440	333
$\theta_1[V^{-1}]$	-1.88	-1.26	1.3	0
$\theta_2\ [V^{-2}]$	2.16	1.76	0	0.92

Table 3-9: Extracted parameters for mobility compact model, using first and second order, depending on the extraction range.

It should be noted that even if the fitting accuracy is acceptable, respective values of μ_0, θ_1 and θ_2 strongly depend on the model used. It emphasizes the fact that these parameters are only fitting parameters. Thus removing either θ_1 or θ_2 would make the model extraction more robust without making the parameter less meaningful. It should be noted that the limited range used for extraction also induces a correlation between μ_0 and θ parameters. Figure 3-18 show the correlation plot between μ_0 and θ_2, extracted with 20 different data samples. A correlation is observed between them.

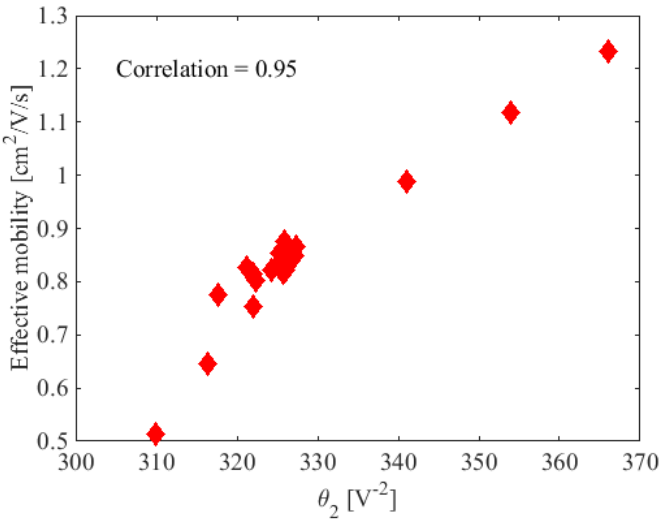


Figure 3-18: Correlation plot of μ_0 and θ_2, extracted using different data samples.

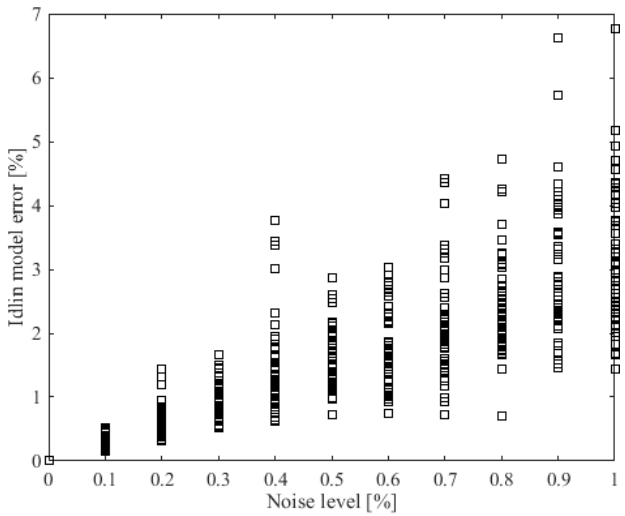


Figure 3-19: RMS error against artificial noise induced in the synthetized data.

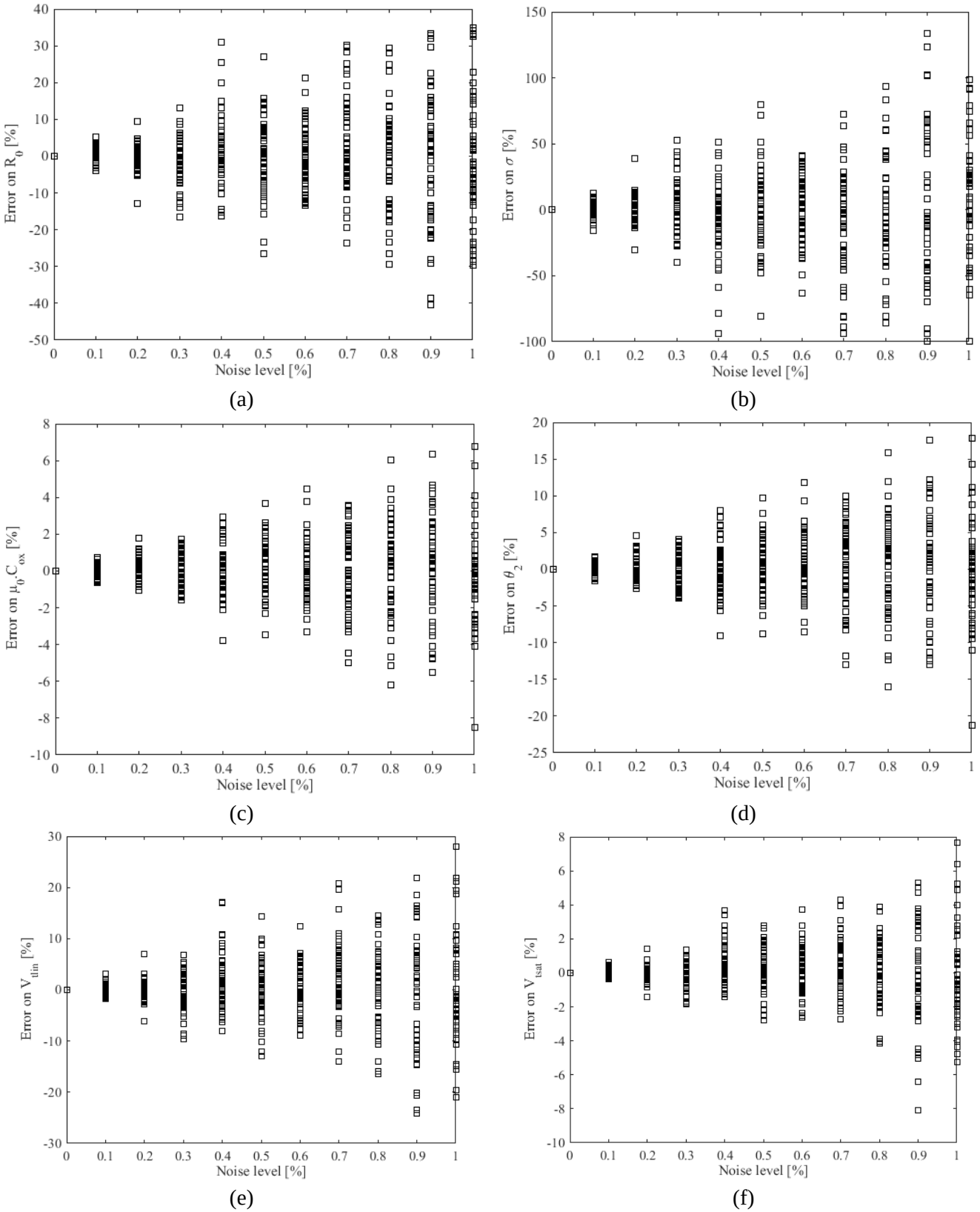


Figure 3-20: RMS error on (a) R_0, (b) sigma, (c) $\mu_0.C_{ox}$, (d) θ_2 (e) Vt_{lin}, (f) Vt_{sat} against artificial noise level.

Figure 3-12 show that θ_1 has less impact on the model than θ_2. Removing θ_1 in the model, the effect of noise on extraction accuracy using synthetic data has been investigated and shows reduced impact on noise compared to the model with θ_1 and θ_2 as shown in Figure 3-19, Figure 3-20 and Figure 3-21.

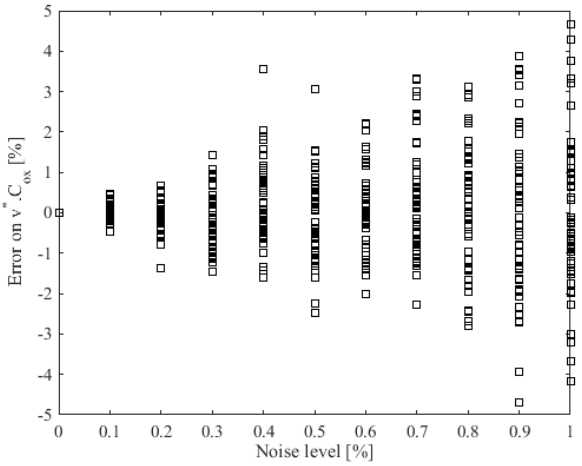
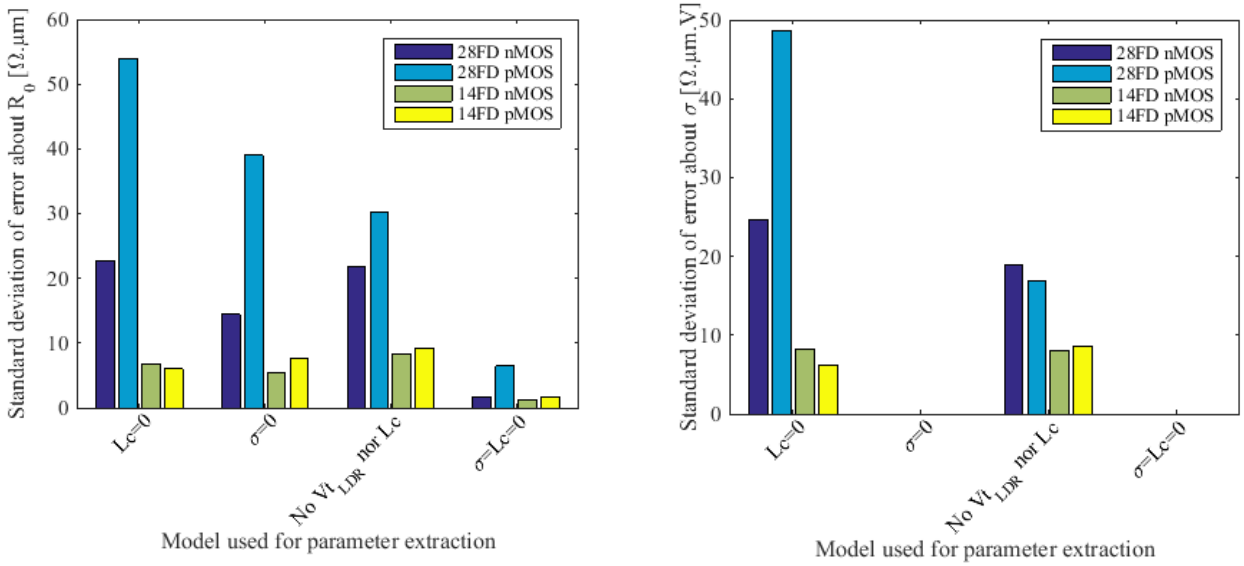


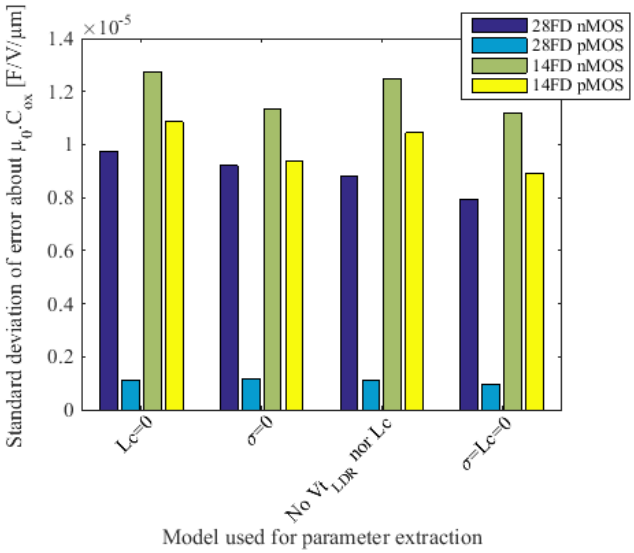
Figure 3-21: RMS error on $v^*.C_{ox}$ against artificial noise.

The error made on the modeled drain current remains now below 7%. The improvement brought by removing θ_1 in the model is important. This emphasizes the need to remove θ_1. Error on extracted model parameter is also well controlled.

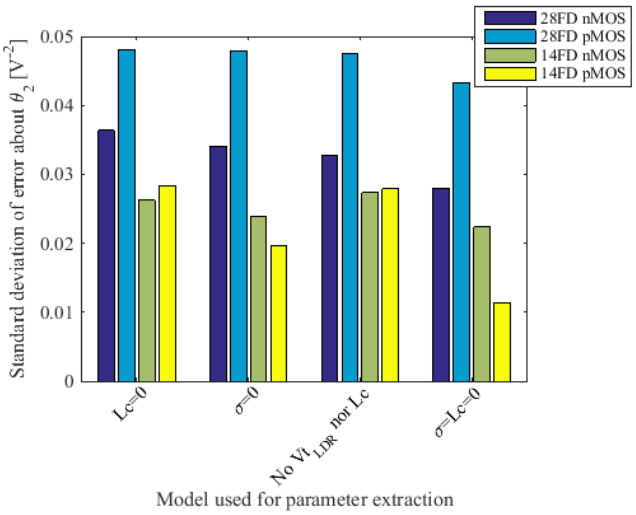
In order to generalize these results, we have run the same test considering model parameters extracted on all technologies and considering four different models. These models are those considered in previous paragraph. Errors on model parameters are gathered in Figure 3-22. In these plots, we show the standard deviation of the error. Considered noise level is set to 1% and θ_1 is set to 0.



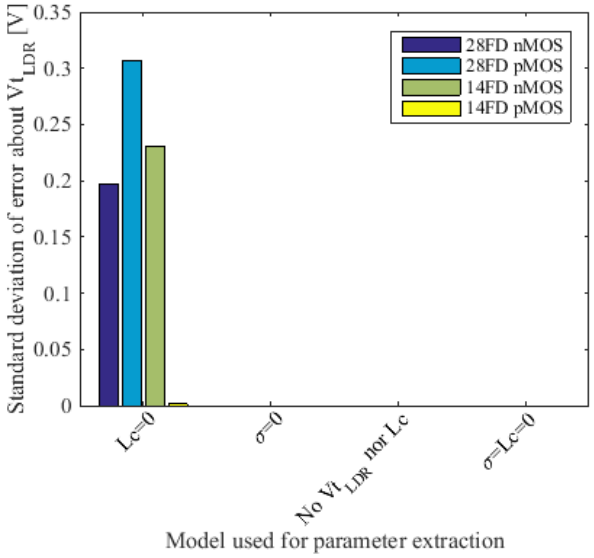
(a) (b)



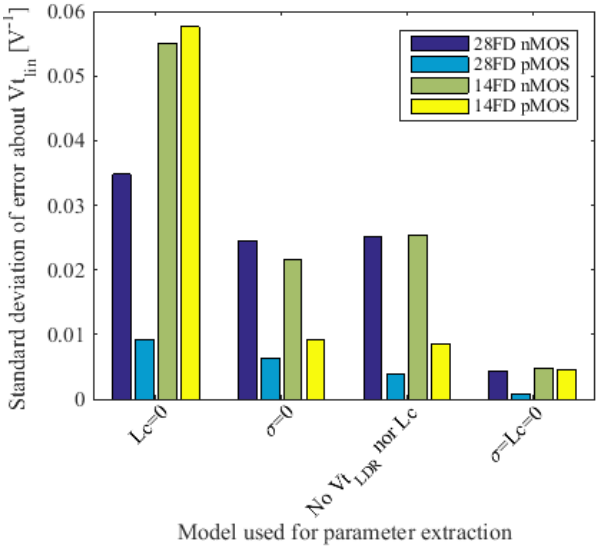
(c)



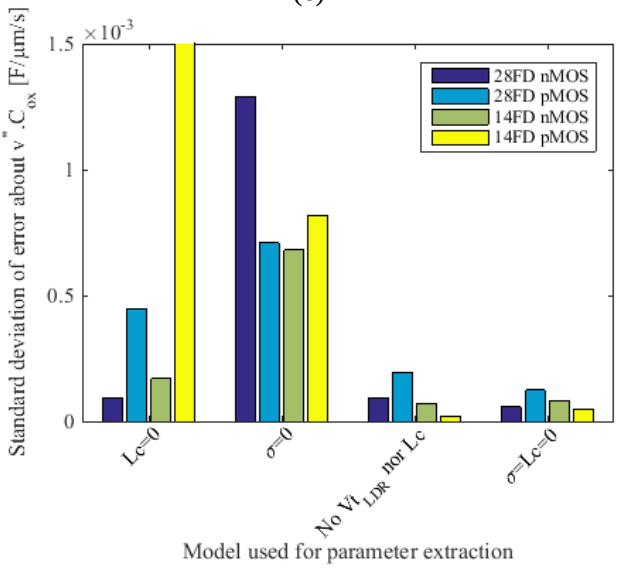
(d)



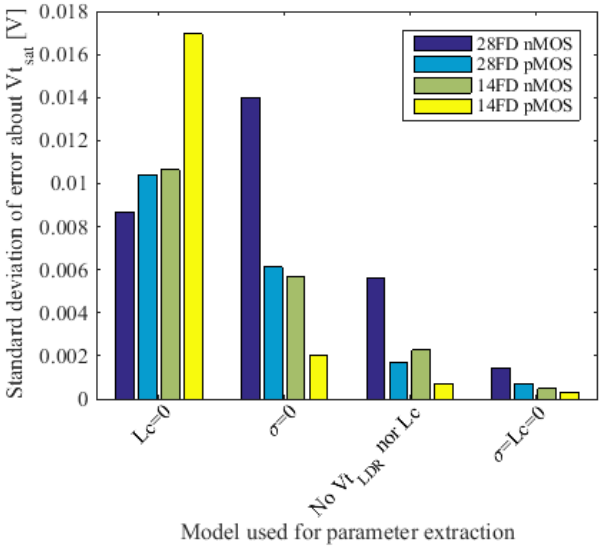
(e)



(f)



(g)



(h)

Figure 3-22: Standard deviation of error on extracted model parameters with 1% noise in measurements, depending on the model and technology considered.

Results show that generally speaking, the more complex the model is, the noisier results are. Serious issues arises, when considering the first model (Lc=0) used to extract 14 nm pMOS FD-SOI model parameters. In this case, $v^*.C_{ox}$ extraction diverges and yields unphysical results due to the noise. Otherwise, noise results remain reasonable.

3.3.4 Conclusion about model parameter extraction method

The study of extraction robustness against synthetic data has revealed that the full model proposed in previous chapter cannot be used as is, considering redundant parameters and the limited amount of data measured in line. Redundant parameters like Vt_{LDR} and Vt_{lin} or Lc and σ have been found to be tricky to extract.

It has been shown on silicon model that fixing these parameters leads to a minimum increase of the model error on silicon extraction. This extraction test has been run considering model parameters extracted on full I_D-V_G of nMOS and pMOS devices of 28 and 14 nm FD-SOI technologies. It has revealed that the model cannot be extracted if all parameters are considered. However as soon as a one parameter is removed, the extraction works fine. An exception must be mentioned for 28 nm pMOS FD-SOI technology where model parameters are badly extracted if only Lc is fixed.

Following that study the effect of artificial noise has been investigated. It revealed that, a small amount of noise can lead to strong error in model extraction. TCAD investigation of the mobility compact model showed that using both θ_1 and θ_2 in the model can lead to a high uncertainty about extraction results. Removing θ_1 allow more robust extractions against noise without making the parameter less meaningful. Noise test has been conducted considering model parameters extracted on full I_D-V_G of nMOS and pMOS devices of 28 and 14 nm FD-SOI technologies and setting θ_1 to 0. Results showed reasonable level of noise in extracted model parameters considering 1% of noise in electrical parameters.

To sum up, attention must be paid to the model used for extraction. If available data for extraction is those of the silicon data sample, we first suggest setting $\theta_1$to 0 in order to reduce the impact of noise in measurements. Then, depending on the device, one or two parameters must be removed. In order to verify the validity of such simplifications, extraction results must be checked. Extraction robustness can be assessed by running the sample size and range test with synthetique data, as it has been done in §3.3.2.1. Then, performing extraction on TCAD simulations, the physical coherence of the results will be checked against the process variations. This will be done in next paragraph. Considering silicon extraction, since many dies are extracted on the same wafer, correlation plots will be performed. Uncorrelated extracted parameters ensure the robustness of the extraction and enable drawing inferences of model parameter's variation impact on drain current. This approach will be used in Chapter 4.

Finally, in order to improve the robustness of the approach, we recommand using a larger data sample size, using more gate biases. This will reduce the effect of noise measurements.

3.4 Application to TCAD simulations

In this paragraph, extractions have been applied on TCAD simulated I_{Dlin} and I_{Dsat} where different process variations have been considered. Results are discussed. The aim of this study is to illustrate, using simulations, how a process variation impacts model parameters. We will demonstrate that model parameters variations can be explained by their physical meaning.

3.4.1 Simulation setup and DOE presentation

TCAD simulations deck is calibrated on 28 nm FD-SOI MOS technology provided by STMicroelectronics. The geometry of the Process Of Reference (POR) is shown in Figure 3-23.

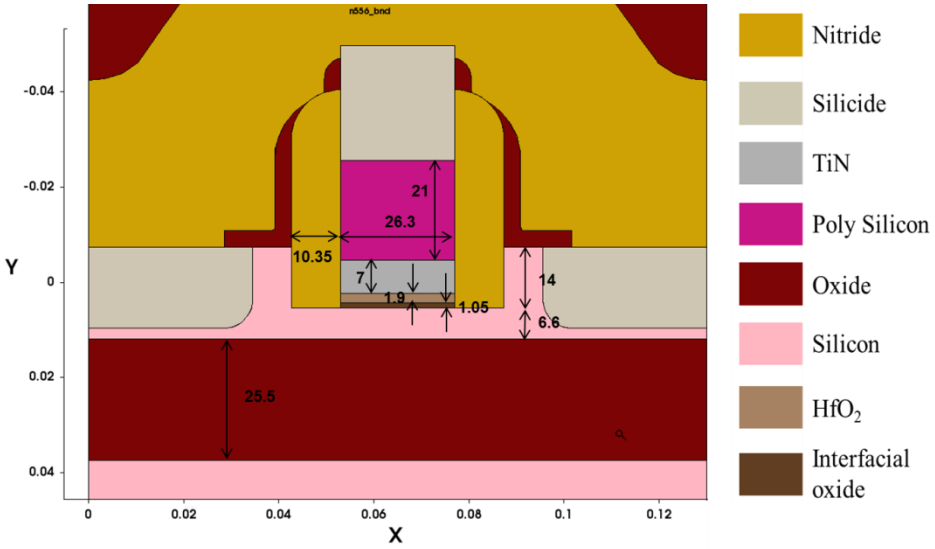


Figure 3-23: Simulated structure using TCAD tools with dimensions in nm

The doping profile for pMOS and nMOS devices is plotted in Figure 3-24. In this plot we see that nMOS device is rather underlapped in contrast with pMOS that is overlapped. We will see later the consequence on extracted model parameters.

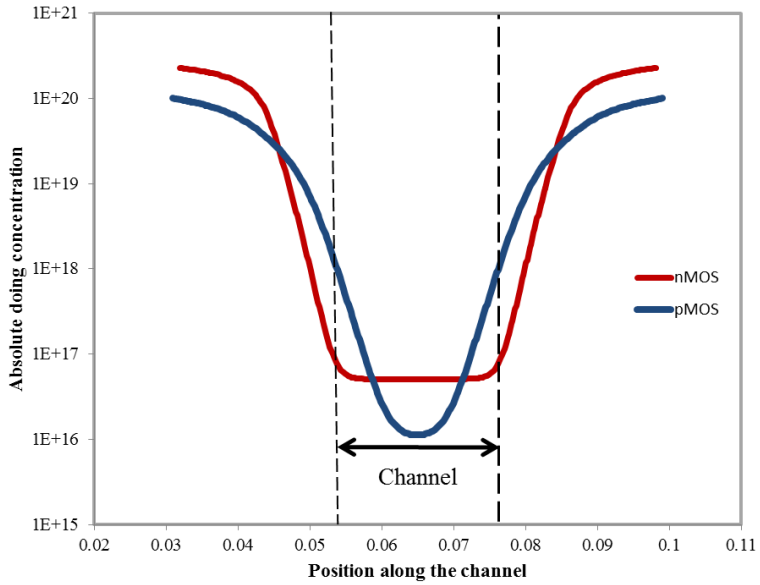


Figure 3-24: Absolute doping level depending on the position along the current path for nMOS and pMOS simulated devices.

Based on this geometry, other simulations have been run with small process parameters variations including:

- Raised source drain epitaxial height (Tepi) [12, 14, 16] nm
- SOI thickness (Tsi) [5, 6, 6.6, 8] nm
- Spacer width (Wsp) [8, 10.35, 12] nm
- Implanted dopant dose (f dose) [0.5, 0.7, 1, 1.2, 1.5] (All source-drain and LDD implant are multiplied by this factor)

- Interfacial layer (IL) thickness (Til) [0.8, 1.05, 1.2, 1.8, 2.5, 4] nm
- IL/High K interfacial charges (Qhk) [10^{10}, 10^{11}, 10^{12}, 3.10^{12}, 10^{13}] cm^{-2}
- Contact resistance (Rext) [20, reference, 200, 500, 2000] Ω
  (Reference values are 90 and 212 Ω for nMOS and pMOS respectively)
- Spike anneal (Tspike) [800, 1000, 1052, 1100]

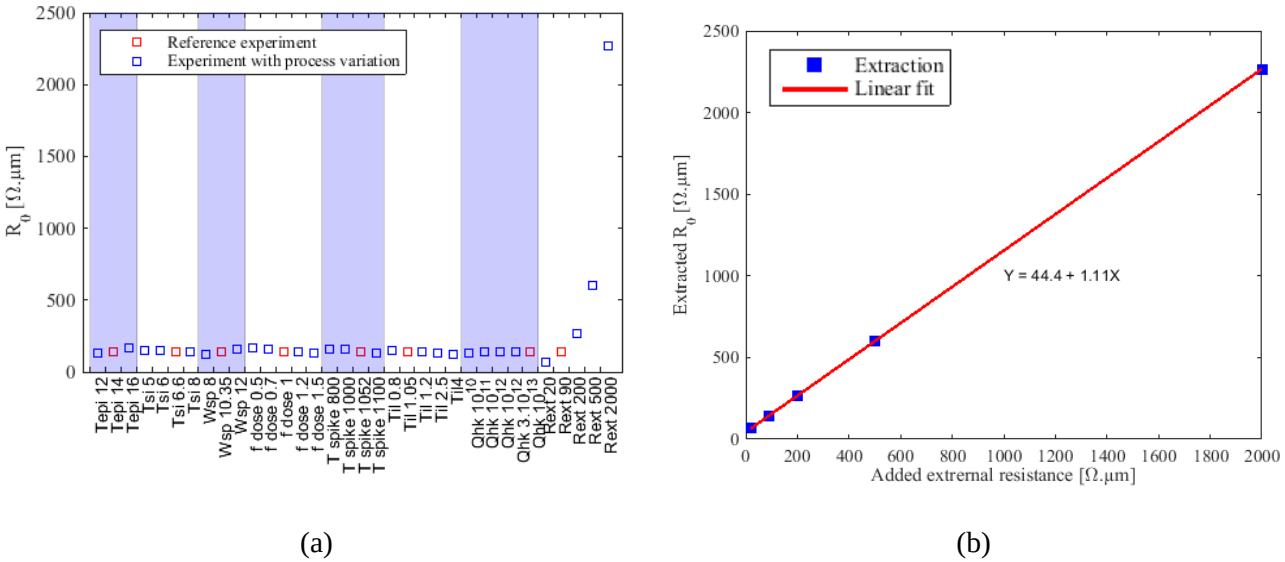
TCAD data sampling used for extraction is gathered in Table 3-5 and Table 3-6 in §3.3.1. This choice of bias conditions is based on available silicon data, in order to keep coherence between both silicon measurements and simulations conclusions. In these simulations, Philips unified model proposed by Klaassen [142] is used for the mobility in combination with high field saturation and thin layer Lombardi model [143]. Neither ballistic transport nor velocity overshoot is simulated here. However mobility model accounts for velocity saturation.

3.4.2 Influence of process variation on extracted model parameters

In this paragraph we present the result of the extraction routine for all experiments of the DOE presented in §3.4.1. We will show that process related model parameters variations are expected based on physical reasoning. Then we will be able to quantify these dependencies using extraction results.

3.4.2.1 Case of nMOS

For the case of nMOS devices Vt_{LDR}, Lc and θ_1 are not considered for data extraction. This model has been chosen for extraction since it yields the most physically coherent results regarding process variations. Using data sample exposed in Table 3-5 and Table 3-6, extraction has been performed on simulated nMOS devices drain current. Results of R_0 extraction are gathered in Figure 3-25 for each experiment of the DOE. Extraction has been performed using silicon data sample. Red dots are the reference experiments. Blue dots are simulated experiments with one process variation with respect to the reference process flow. White and shaded strips gather experiment according to their common variable process parameter.



(a) (b)

Figure 3-25: (a) R_0 extracted on TCAD simulated I_D-V_G including experiment with process variations. (b) Extracted R_0 against added external resistances on TCAD simulations.

In Figure 3-25 (a), we see that, as expected, R_0 is mostly sensitive to external resistance. Quantitative variations of R_0 against added external resistance are shown in Figure 3-25 (b). It is shown that

extracted parameters track well the implemented one. From the linear extrapolation of that scatter plot, we can deduce the highly doped source-drain region resistance that is 44.4 $\Omega.\mu m$. Results of σ extraction depending on the process variation of the DOE presented beforehand are gathered in Figure 3-26.

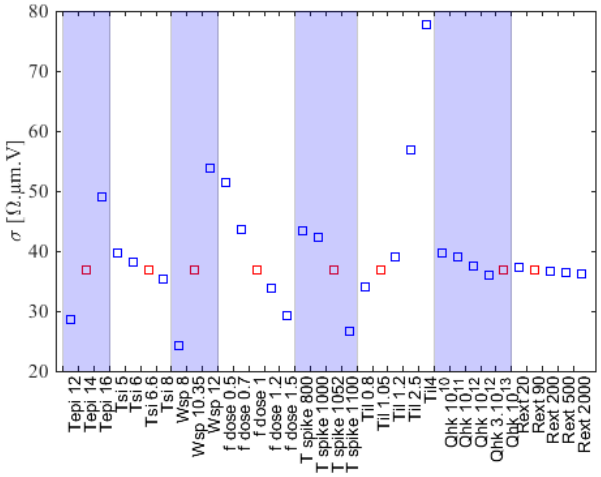
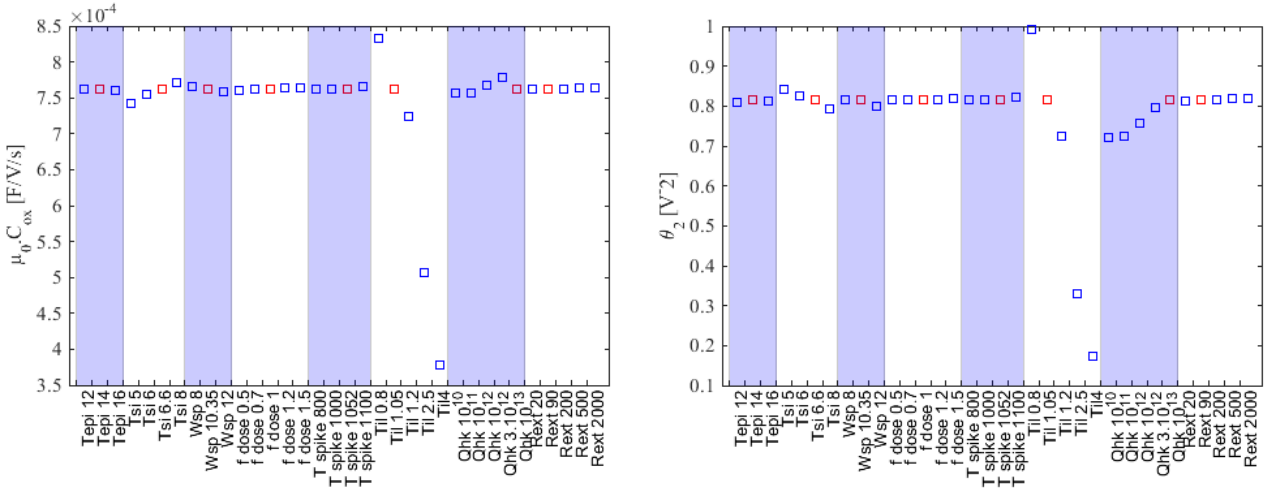


Figure 3-26: σ extracted on TCAD simulated I_D-V_G including experiment with process variations.

σ parameter is related to the V_G dependent part of the access resistance. It is shown in that plot that σ depends on every parameter except the external resistance. In chapter 2 we have seen that gate voltage dependent access resistance depends on LDR doping level. Consequently, the more the transistor is underlapped, the greater σ. Thus when Tepi or Wsp is large, the junction is moved away from the gate. The transistor becomes underlapped and σ rises. Tsi influences σ as well. The dopant dose used for implant (f dose) acts directly on the LDR doping concentration thus the lower the dose, the higher σ. Thick T_{il} reduces the field from the gate. Thus the higher is T_{il}, the higher is σ. When the anneal temperature is increased, dopants migrate farther. Thus LDR becomes more doped and σ diminishes. However σ does not depend on external resistance since this resistance does not impact the LDR. This enphasises the robustness of the extraction and is a validation about σ physical interpretation. It validates its implementation.

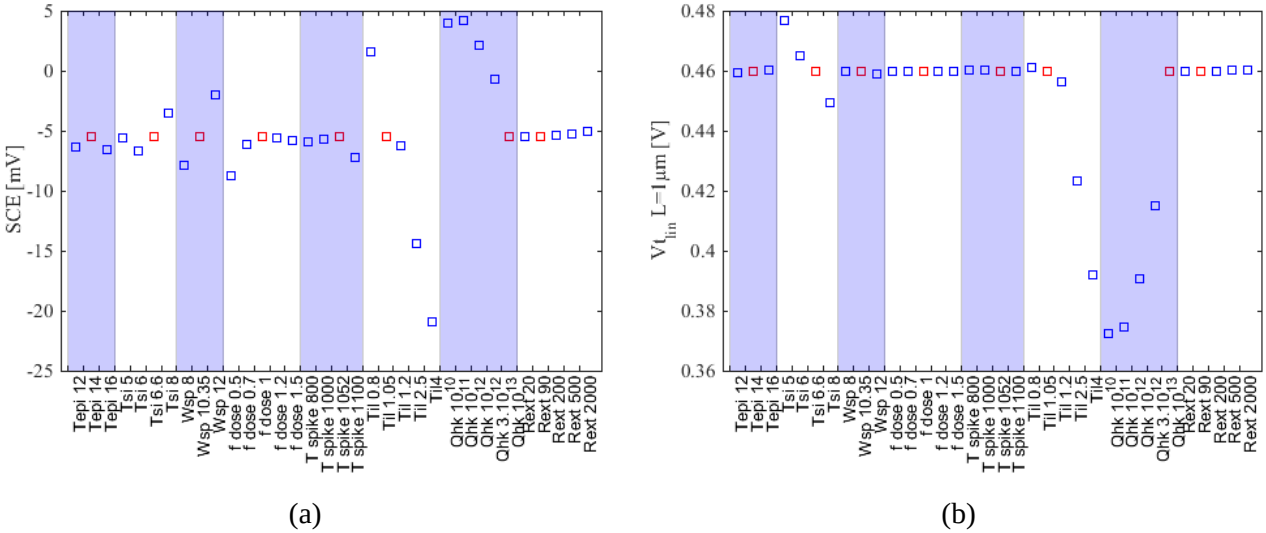


(a) (b)

Figure 3-27: $\mu_0.C_{ox}$ (a) and θ_2 (b) extracted on TCAD simulated I_D-V_G including experiment with process variations.

Figure 3-27 presents $\mu_0.C_{ox}$ and θ_2 extracted for all the simulated experiments. As it has been discussed in Chapter 2, μ_0 and θ_2 are complex functions that depend on C_{ox}, C_{si} (and thus on V_B and Tbox). Consequently $\mu_0.C_{ox}$ and θ_2 depend on Tsi and T_{il}. In addition θ_2 and $\mu_0.C_{ox}$ depend on Qhk that are the charged defect at the high K interface. This is the effect of remote Coulomb that could be captured through μ_0 and/or θ_2. The most influent process parameter is T_{il}, that is inversely proportional to C_{ox}, thus to $\mu_0.C_{ox}$.

Figure 3-28 (a) and Figure 3-28 (b) represent Vt_{lin} for short and long channel transistors respectively. These two parameters essentially depend on channel related process parameters (Tsi, T_{il} and Qhk). In contrast with long channel Vt_{lin}, short channel Vt_{lin} is slightly impacted by access related process parameters. Indeed, since the access plays a major role in the current drain characteristics of short channel devices, it induces parasitic effects in the extraction.



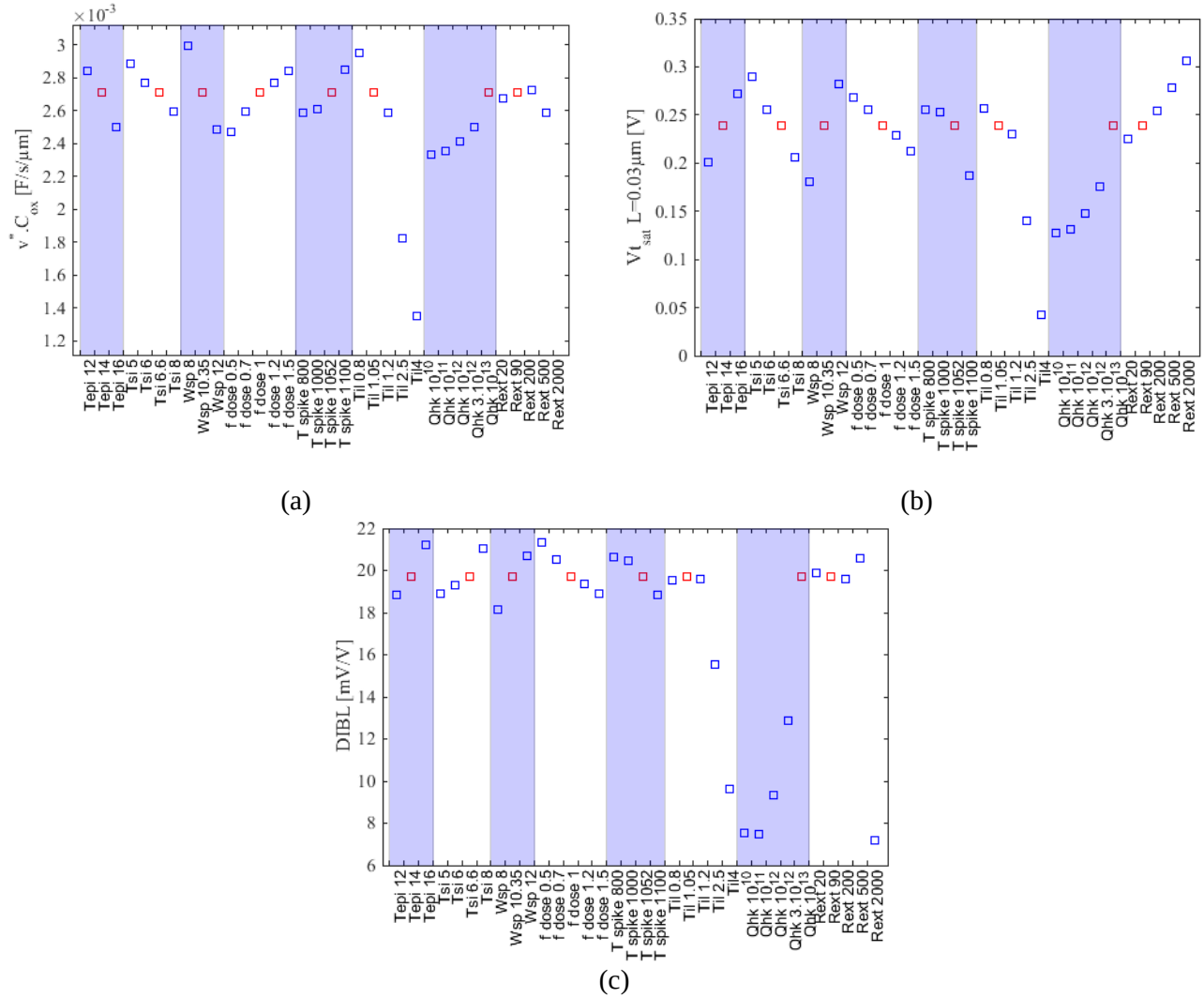
(a) (b)

Figure 3-28: Vt_{lin} for short (a) and long (b) channel devices extracted on TCAD simulation against process variations.

Figure 3-37 shows Vt_{lin} against process variations for long channel devices as well as Short Channel Effect (SCE) coefficient. SCE coefficient is calculated according to the following equation.

$$V_{tlin\text{-}short} - V_{tlin\text{-}long} = \Delta V_t = SCE \quad \textbf{(121)}$$

SCE reflects the loss of electrostatic control on the channel with decreasing gate length. It depends on the source-drain junction position. If the transistor is overlapped, electrostatic control will be lost more quickly with decreasing gate length. Indeed in this case, the distance between the junctions is shorter. Thus, considering the same gate length, an overlapped transistor will have less electrostatic control than a transistor with junctions well aligned with the gate. As a consequence, narrow spacers induce greater negative SCE by overlapping the transistor. SCE also depends on Til and and Qhk. Indeed, a thick Til reduces the gate electrostatic field in the channel, thus the electrostatic control of the channel is weaker. Interfacial charges Qhk shield the gate electrostatic and increase SCE as well.



(a) (b)

(c)

Figure 3-29: $\mathbf{v_{sat}.C_{ox}}$ (**a**) **and** $\mathbf{Vt_{sat}}$ (**b**) extract**ed on TCAD simulated** $\mathbf{I_D}$-$\mathbf{V_G}$ **including experiment with process variations.**

Figure 3-29 (a), (b) and (c) show the variations of $v^*.C_{ox}$, Vt_{sat} and Drain Induced Barrier Lowering (DIBL) respectively. DIBL has been calculated following:

$$DIBL = \frac{Vt_{lin} - Vt_{sat}}{V_{dd} - Vd_{lin}} \tag{122}$$

$v^*.C_{ox}$ should only depend on T_{il} through C_{ox} parameter. However it is also slightly sensitive to other parameters. This inconsistency could be due to the fact the self-heating is not taken into account in our model while it is simulated in TCAD. The same consideration holds for Vt_{sat} and DIBL. $v^*.C_{ox}$ could not be extracted when R_{ext} is greater than 500 $\Omega.\mu m$. Indeed when $R_{ext} = 2000\,\Omega.\mu m$, the external resistance drives the saturation drain current and v^* has no significant impact anymore and cannot be extracted properly.

In order to investigate the self heating effect, POR has been simulated with and without self-heating. Extraction has then been performed for both cases. The difference between model parameters without and with self heat has been calculated. Results shown in Figure 3-30 reveal that self-heating mostly reduces $v^*.C_{ox}$ and Vt_{sat} by a non-negligible amount. Thus self-heating does not impact linear model parameters extraction but only saturation model parameters.

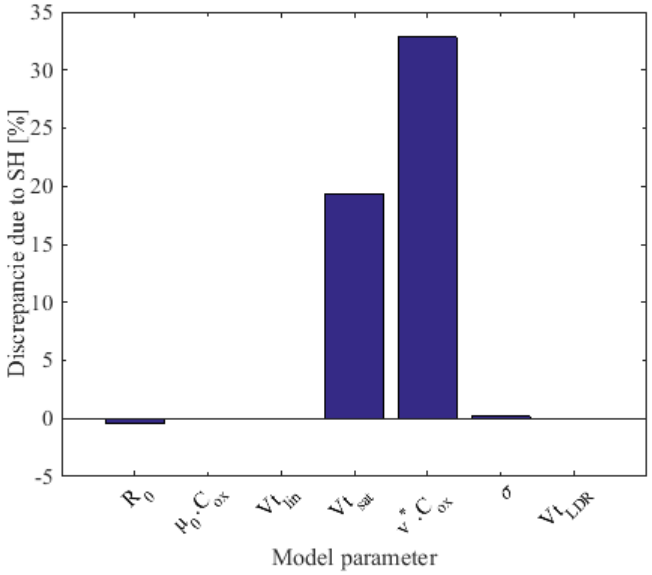
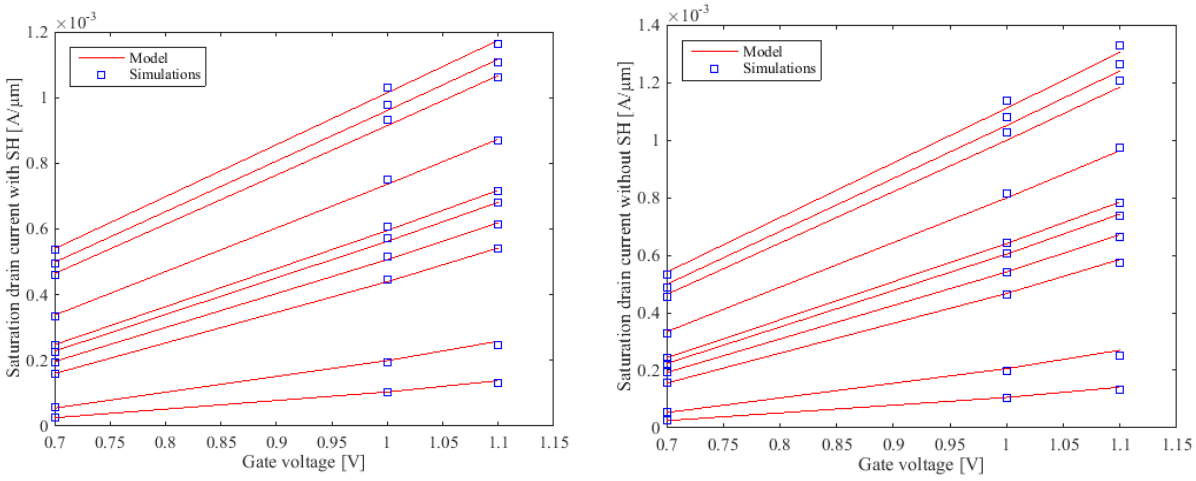


Figure 3-30: Discrepancies between model parameters extracted with and without self-heating (SH)

Figure 3-31 shows modeled and simulated saturation drain current against V_G for different gate length with and without self heating. We see that self-heating tends to reduce drain current at high V_G. In presence of self-heating, the second derivative of I_{Dsat} with respect to V_G is negative but the model cannot account for such a behavior. Thus self-heating effect is accounted for though $v^*.C_{ox}$ and V_{tsat} parameters.

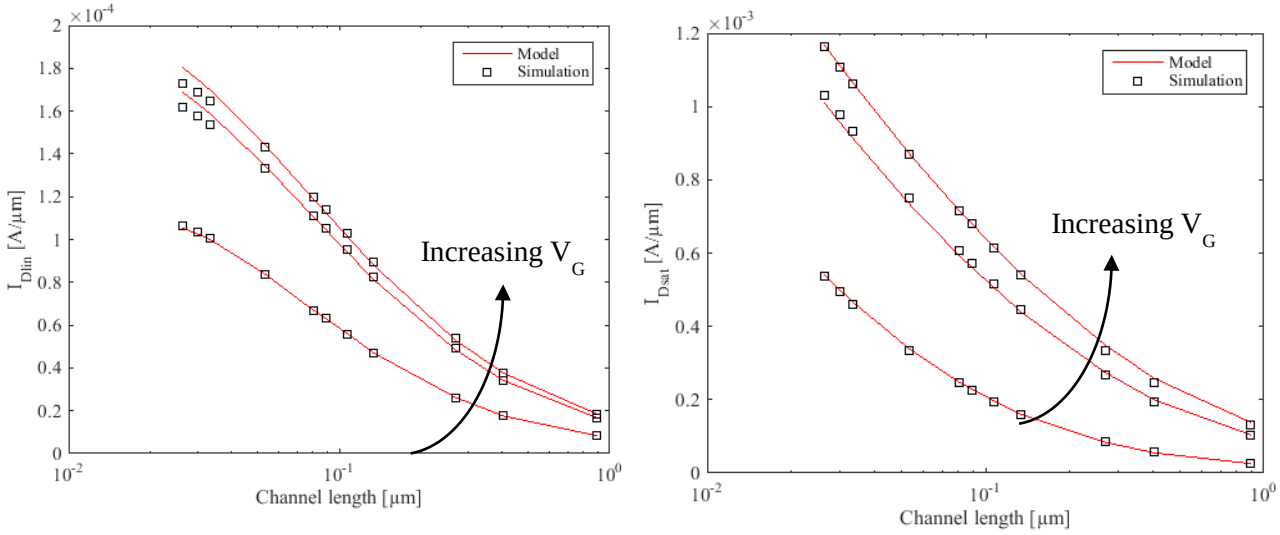


(a) (b)

Figure 3-31: Saturation drain current simulated and modeled with (a) and without (b) self-heating (SH).

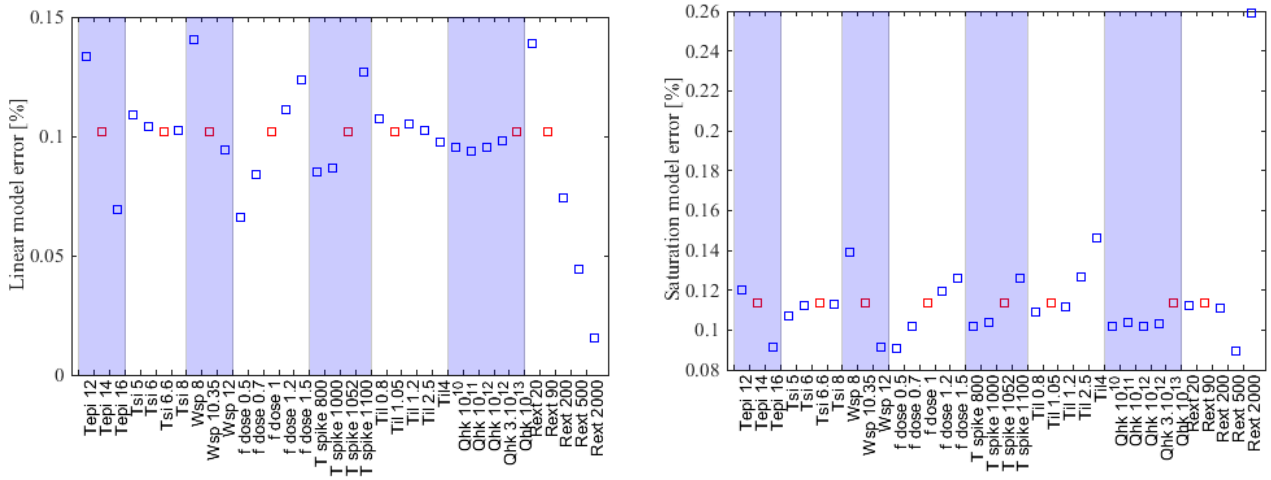
Figure 3-32 (a) and Figure 3-32 (b) show simulated and modeled linear and saturation drain current respectively. A good fit is obtained. Linear drain current is slightly overestimated for short channel devices at high gate voltage. This is due to the simplification of the model (no Vt_{LDR}, no Lc and no θ_1).

Figure 3-33 plots the relative error made by the model depending on the experiment considered. Even though the error is experiment dependent, it remains relatively low, assessing the robustness of the approach.



(a) (b)

Figure 3-32: Linear (a) and saturation (b) drain current modeled and simulated against L for different V_G.



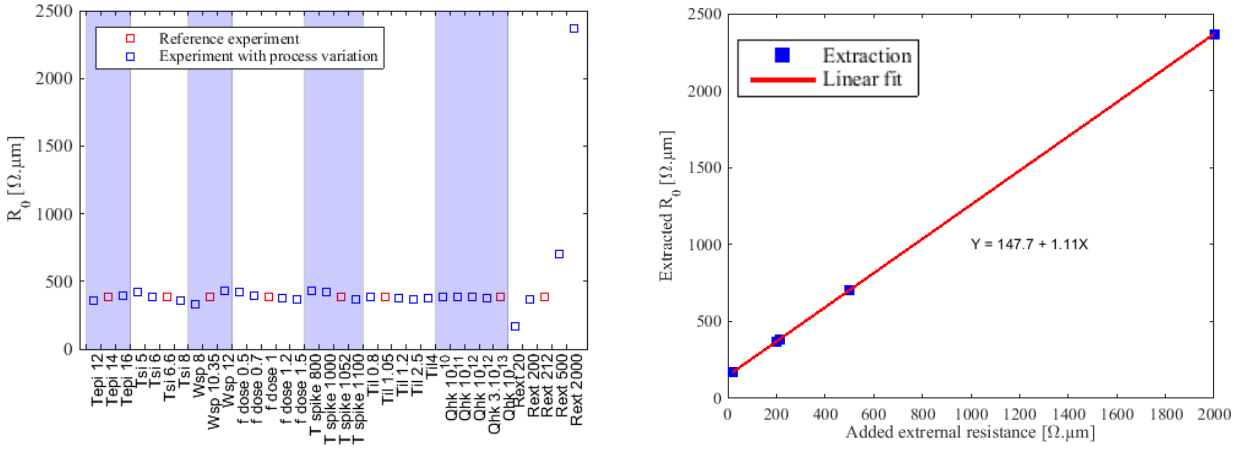
(a) (b)

Figure 3-33: Model error on linear (a) and saturation (b) drain current

3.4.2.2 Case of pMOS

In this paragraph we expose the results of model parameters extraction on pMOS devices. The main difference between nMOS and pMOS devices is the carrier mobility (that is lower for pMOS since holes effective mass is greater than electrons) and the doping profile (for our case); see Figure 3-24. We will see how these differences affect model parameters.

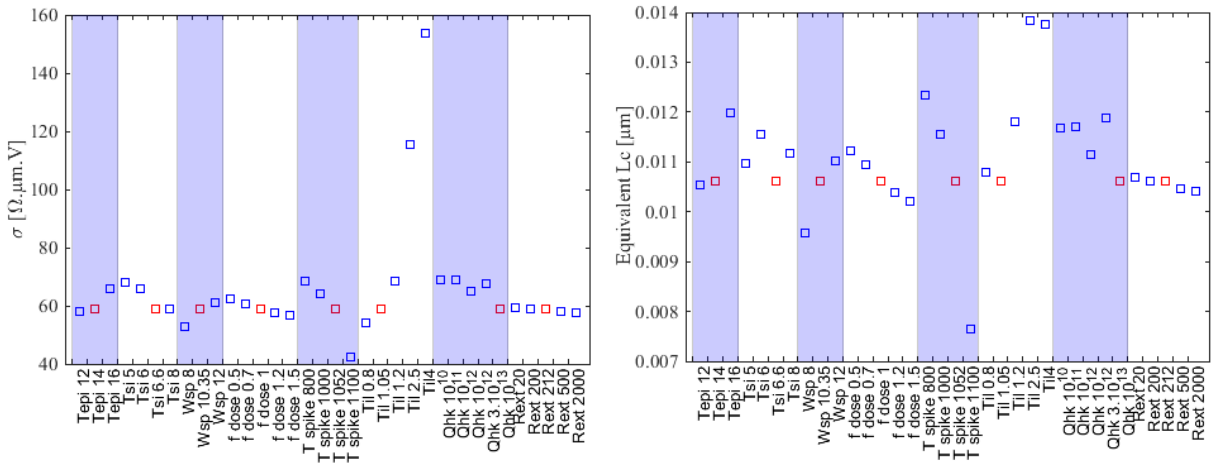
The equation used to model pMOS TCAD simulation is the same than nMOS. Thus Vt_{LDR}, Lc and θ_1 are not considered for data extraction. Using data sample exposed in Table 3-5 and Table 3-6, extraction has been performed on simulated pMOS devices drain current. First of all, Figure 3-34 (a) shows the extracted R_0. Again here we see that R_{ext} drives R_0 value, and the highly doped source-drain resistance can be extrapolated from R_0 against R_{ext} plot. This resistance is much higher than the one of nMOS devices (147.7 $\Omega.\mu m$ compared to 44.4 $\Omega.\mu m$). Indeed we can see from Figure 3-24 that doping concentration in this region is lower in pMOS than nMOS ($10^{20}\ cm^{-3}$ for pMOS compared to $2.3 \cdot 10^{20}\ cm^{-3}$ for nMOS). Moreover, hole mobility is lower than electron ones.



(a) (b)

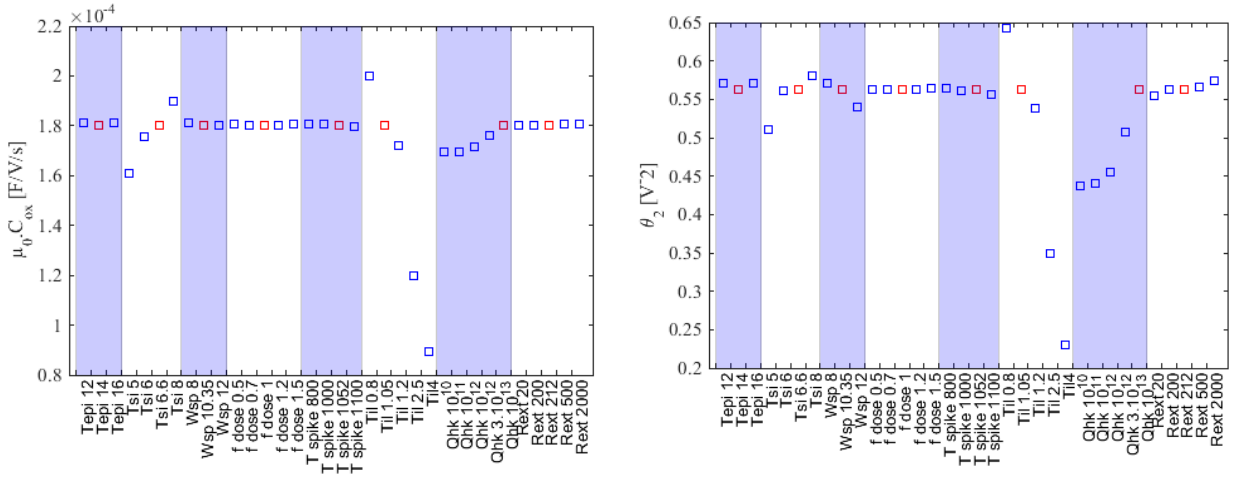
Figure 3-34: (a) R_0 extracted on TCAD simulated I_D-V_G including experiment with process variations. (b) Extracted R_0 against added external resistances on TCAD simulations.

R_0 is also sensitive to some other parameters. Increasing implant dose or increasing the annealing temperature reduces R_0 since it increases the amount of active dopants in the highly doped source-drain regions. Increasing Tepi or Wsp increases R_0 since it lengthen the path from the silicide contact point to the entry of the channel.



(a) (b)

Figure 3-35: σ and equivalent Lc extracted on TCAD simulated I_D-V_G including experiment with process variations.

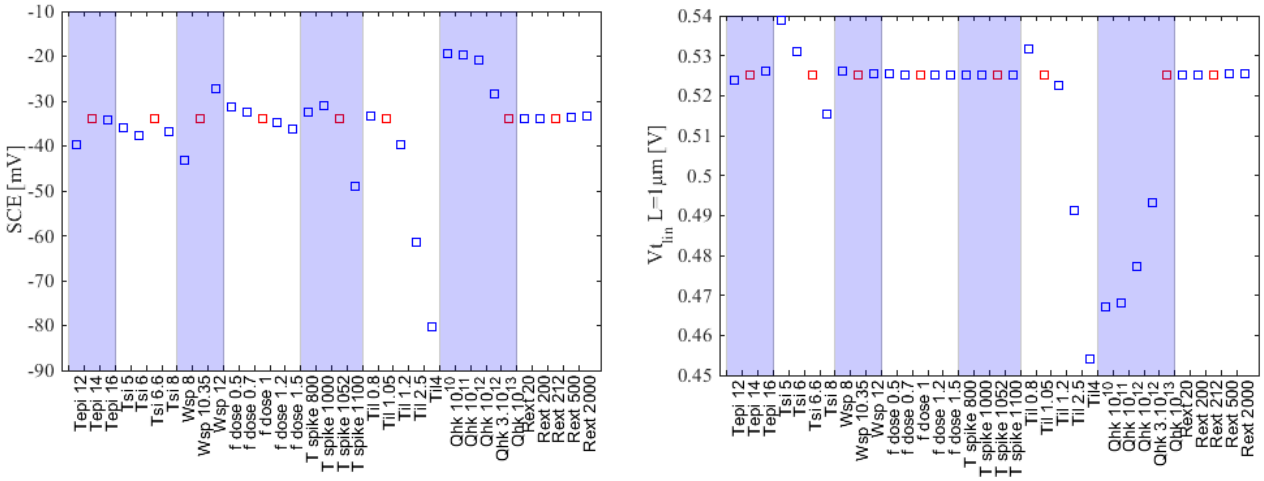


(a) (b)

Figure 3-36: $\mu_0.C_{ox}$ (a) and θ_2 (b) extracted on TCAD simulated I_D-V_G including experiment with process variations.

Figure 3-35 (a) shows extracted σ against process variations. pMOS is much more overlapped that nMOS, thus LDR is more sensitive to the gate properties. This explains the strong dependence on T_{il} compared to other parameters. However R_{ext} does not influence σ value as expected. As it has been explained previously, in linear regime, there is a perfect equivalence between σ and Lc if $V_{tLDR} = V_{tlin} + \frac{V_D}{2}$.

$\mu_0.C_{ox}$ and θ_2 variations are shown in Figure 3-36. Process dependence is similar to the case of nMOS: T_{il}, Tsi and Qhk are the only parameters that impact θ_2 and $\mu_0.C_{ox}$ as expected.

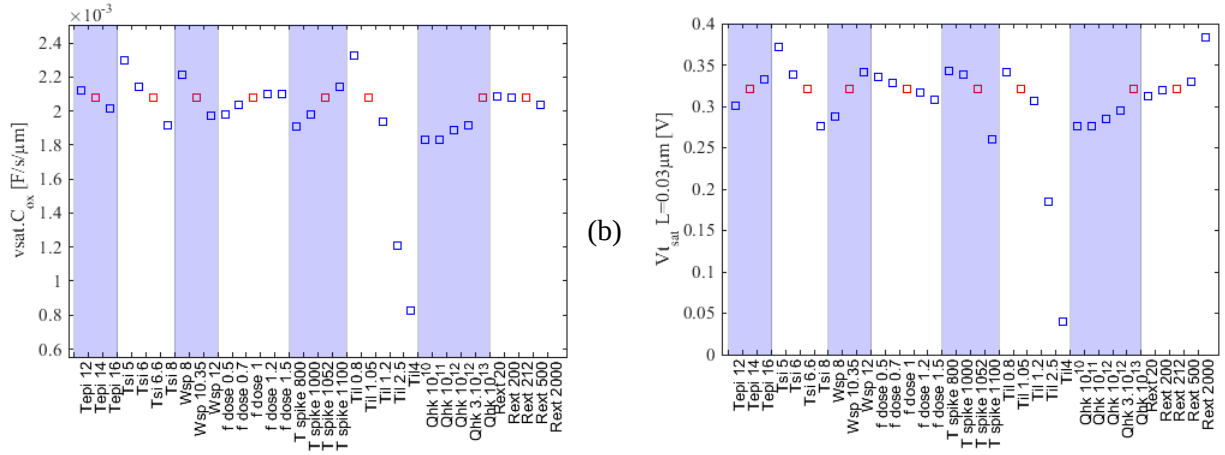


(a) (b)

Figure 3-37: Vt_{lin} for short (a) and long (b) channel devices extracted on TCAD simulation against process variations.

Figure 3-37 shows Vt_{lin} against process variation for long channel devices as well as SCE coefficient. SCE is weaker in nMOS ($\sim 5.10^{-3}$ V.µm) than pMOS ($\sim 30.10^{-3}$ V.µm). Indeed, pMOS is more overlapped than nMOS. Considering long channel threshold voltage, channel related process parameters (T_{il}, Tsi and Qhk) are by far the most influent process parameters.

(a) (b)



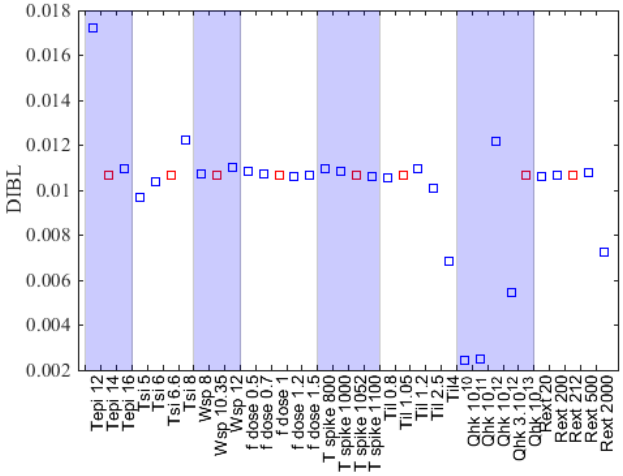


Figure 3-38

shows $v_{sat}.C_{ox}$, Vt_{sat} and the DIBL against process variations. Again, $v_{sat}.C_{ox}$, V_{tsat} and DIBL depends on access parameters due to self-heating that is not accounted for in the model. However T_{il} remains the most influent parameter on $v^*.C_{ox}$ since it has a direct impact on C_{ox}. pMOS is half nMOS one due to different position of the junction.

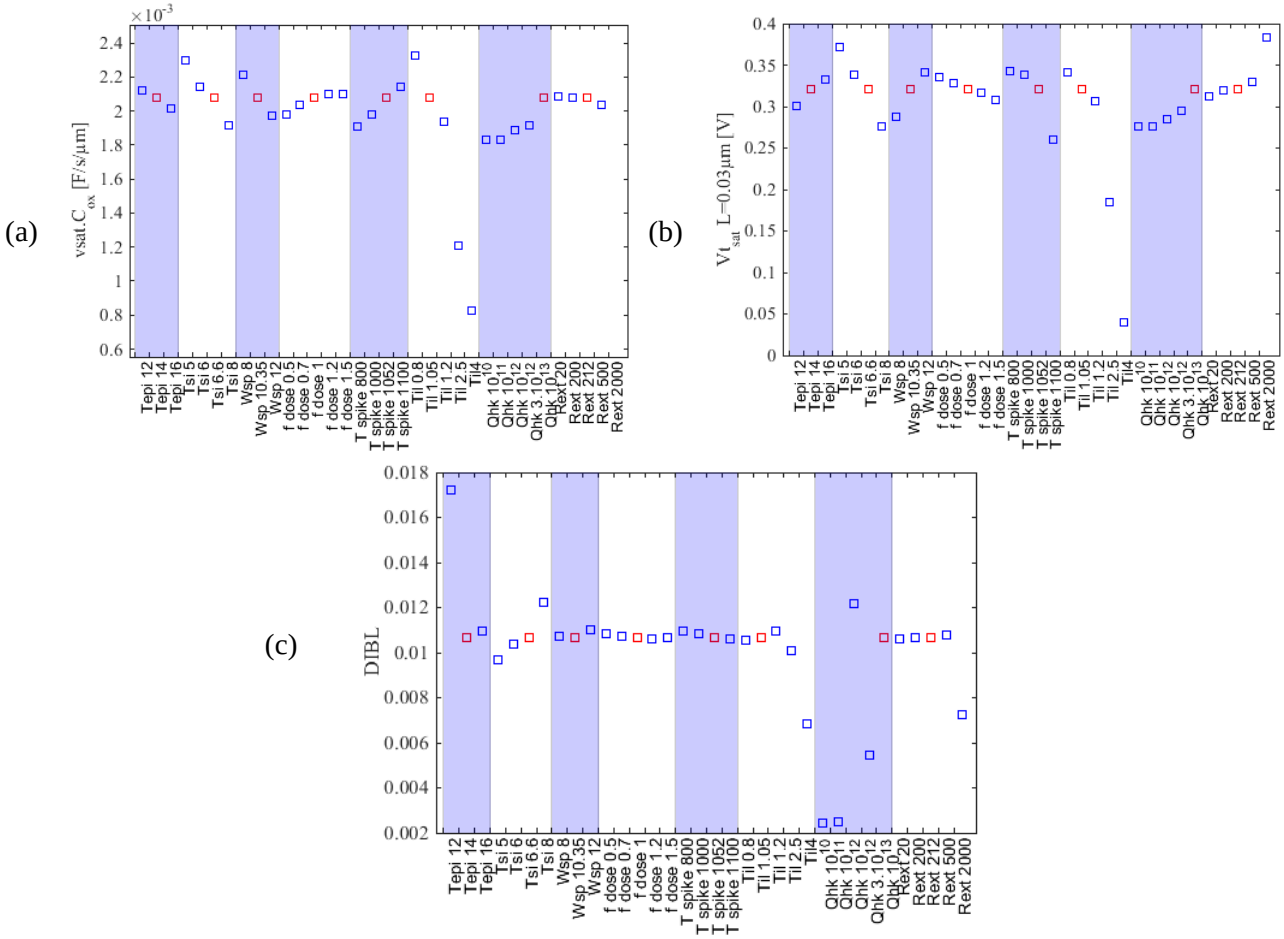


Figure 3-38: $v_{sat}.C_{ox}$ (a), Vt_{sat} (b) and DIBL extracted on TCAD simulated I_D-V_G including experiment with process variations.

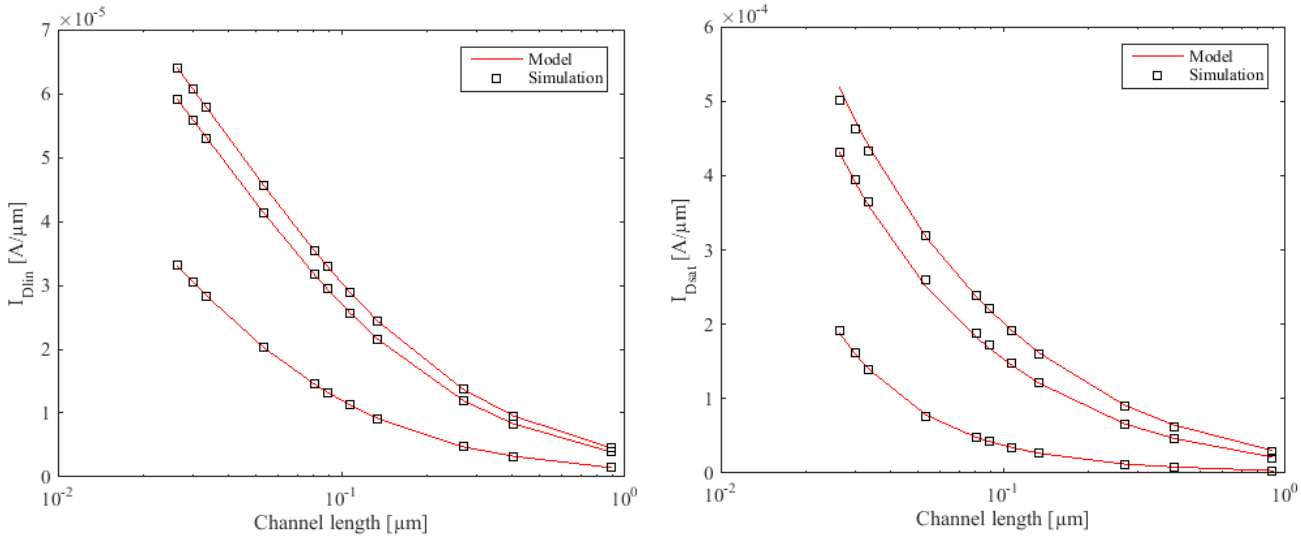


Figure 3-39: pMOS linear drain current modeled and simulated against L for different V_G. Plotted L and V_G are the one used for the extraction procedure.

Figure 3-39 shows modeled and simulated drain currents against channel length for reference experiment. We see a very good fit. The model simplifications does not impact the fitting accuracy, thus Lc and Vt_{LDR} are not influent model parameters in this case.

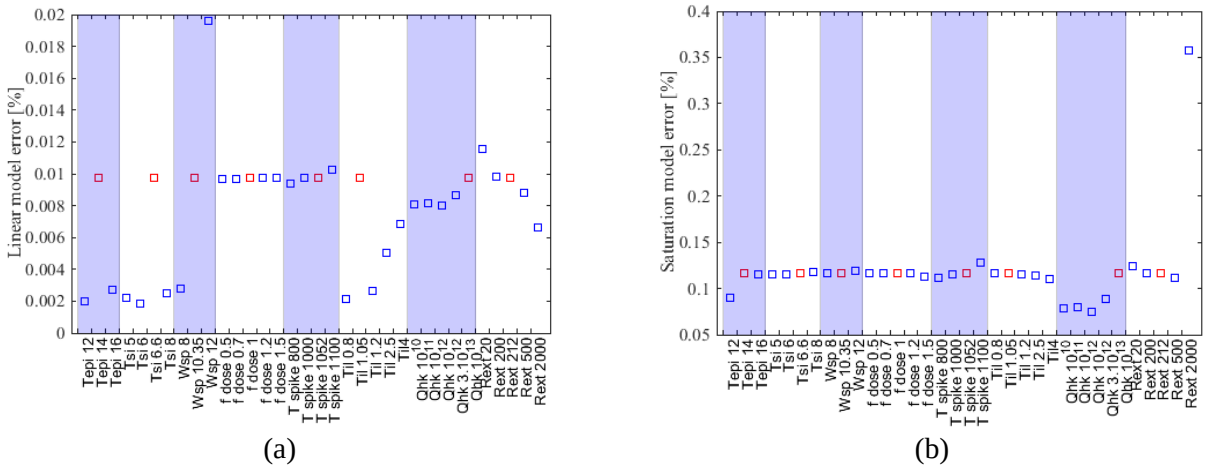


Figure 3-40: Model error on linear (a) and saturation (b) drain current

Figure 3-40 shows the linear and saturation drain current RMS error against process variations. We see that this error remains low no matter the experiment, confirming the robustness of the model and the extraction method.

3.5 Conclusion

Following the introduction of the compact model in chapter 2, a method to extract model parameters has been introduced in this chapter. The method is decomposed into 3 steps. First step consists in extracting linear model parameters using linear least square fit. Then these values are used as a first guess for a nonlinear optimizer that refines parameters value. Finally, saturation model parameters are extracted using nonlinear least square fit.

This method has been tested to assess its robustness before applying it on PT silicon measurements. Tests have been conducted on synthetized data against sample size and range considering model parameters extracted on full I_D-V_G of nMOS and pMOS devices of 28 and 14 nm FD-SOI

technologies. We have seen that data sample ranges and sizes available in silicon measurements are too small to properly extract all model parameters. Removing successively each parameters from the model showed that θ_1, Vt_{LDR} and Lc are the least significant model parameters. Extraction test has been run once again considering cases where some of these parameters have been fixed. It showed that as soon as one parameter is removed, the extraction works fine. An exception must be mentioned for 28 nm pMOS FD-SOI technology where model parameters are badly extracted if Lc is the only fixed parameter. Thus removing one parameter allows robust extraction with a minimum error in the model.

Following that study, the effect of artificial noise has been investigated. It revealed that, a small amount of noise can lead to strong errors in model extraction. TCAD investigation of the mobility compact model showed that using both θ_1 and θ_2 in the model can lead to a high uncertainty about extraction results. Removing θ_1 allows more robust extractions against noise without making the parameter meaningless. Noise test has been conducted considering model parameters extracted on full I_D-V_G of nMOS and pMOS devices of 28 and 14 nm FD-SOI technologies and setting θ_1 to 0. Results showed reasonable level of noise in extracted model parameters considering 1% of noise in electrical parameters.

To sum up, attention must be paid to the model used for extraction. First we suggest setting $\theta_1$to 0 in order to reduce the impact of noise in measurements. Then, depending on the device, one or two parameters must be removed. In order to verify the validity of such simplifications, extraction results must be checked. Considering TCAD simulations, the physical coherence of the results will be checked against the process variations. Considering silicon extraction, since many dies are extracted on the same wafer, correlation plots will be performed. Uncorrelated extracted parameters ensure the robustness of the extraction and enable drawing inferences of model parameter's variation impact on drain current.

The extraction procedure has been run on a TCAD simulated DOE. The DOE account for different process parameters (External resistance, epitaxial thickness, SOI thickness, spacer width, implanted dose, annealing temperature, insulating layer thickness and defects at the high K interface). We have shown that model parameters response to process variations is physically coherent, testifying on model parameters physical meaning and extraction robustness. Extractions have been run for nMOS and pMOS enabling a quantification of the impact of active dopant dose in the source-drain region as well as the junction profile on the drain current and model parameters.

Chapter 4 :
Compact modeling: application to 28 nm and 14 nm FD-SOI technologies

In this chapter, the extraction method developed in previous chapter is applied to 28 and 14 nm FD-SOI silicon devices measurements. Model parameters variations with process variations are investigated and we will see how model extraction helps getting more insights into the device characteristics and help interpreting the relation between process parameters and device performances. The chapter starts with extraction on 28 nm FD-SOI technology in § 4.1. Effect of Dynamic Surface Anneal (DSA) and source drain implant dose and energy are studied. In the same trend, in §4.2, we apply extraction method on 14 nm FD-SOI silicon data. Effect of source-drain dopant concentration, HF clean before epitaxy and epitaxial thickness are investigated. In addition, within-wafer variability is addressed in §4.3. Forward and backward propagation of variance as well as Monte Carlo draws are used to model this variability.

4.1 Application to 28 nm FD-SOI technology

In this paragraph, we show the results of extraction applied to 28 nm FD-SOI devices measurements. Extractions are carried on several wafers with process variations. Details of the DOE and experimental setup is discussed in §4.1.1. In §4.1.2 the extraction accuracy will be assessed and we will see how the extraction enables a clear quantification of the process impact on device characteristics. We will then bring a physical interpretation of the variations and we will see that model parameters variations depending on process variations are well correlated with extraction results based on TCAD simulations.

In order to model the device drain current we use equations developed in chapter 2 where Lc and θ_1 have been set to zero. Contrary to extractions performed on TCAD simulations, here we extract V_{tLDR} parameter as well. We will show that extraction performe well using this model thanks to correlation plots of model parameters. The extraction robustness using this equation will be tested using correlation plots of extracted model parameters. As a reminder, we recall here the model that will be used in further extraction for linear drain current:

$$Id_{lin} = \frac{V_{DS}}{R_{tot}} \tag{123}$$

where the total width normalized transistor resistance R_{tot} is:

$$R_{tot} = R_0 + \frac{\sigma}{V_{GS} - V_{tLDR}} + \frac{L}{\mu_0 . C_{ox}} \left(\frac{1}{\left(V_G - V_t - \frac{V_{DS}}{2}\right)} + \theta_2 \left(V_G - V_t - \frac{V_{DS}}{2}\right) \right) \tag{124}$$

The total resistance is simply the sum of contact and source-drain resistance represented by R_0 term, the LDR resistance and the channel resistance.

Saturation drain current is expressed as:

$$Id_{sat} = \frac{Id'_{sat}}{1 + G_m \cdot R_S} \tag{125}$$

where $R_s = \frac{R_0 + \frac{\sigma}{V_{GS} - Vt_{LDR}}}{2}$, Id'_{sat} is the intrinsic saturation drain current:

$$Id'_{sat} = \frac{W}{L}\mu_{eff}C_{ox}\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)V_{Dsat} \quad (126)$$

G_m is the V_G derivative of Id_{lin}:

$$G_m = 4\frac{W}{L}\mu_{eff}C_{ox}V_{Dsat}\cdot\left(A - \theta_2\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)^2\right) \quad (127)$$

where $A = 1 + \frac{1}{L}\left(\frac{V_{Dsat}\cdot\mu_0}{v^*}\right)$ and μ_{eff} is the effective mobility, accounting for scattering mechanisms, velocity saturation and ballistic transport:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta_2\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)^2 + \frac{V_{Dsat}\cdot\mu_0}{Lv^*}} \quad (128)$$

and V_{Dsat} is the drain saturation voltage and is derived as V_{DS} such as $\frac{dId_{lin}}{dV_{DS}} = 0$:

$$V_{Dsat} = 2\frac{u - \sqrt{u.\left(1 + 2.\frac{\mu_0}{v^*.L}(V_G - V_{tsat})\right)}}{-\frac{2\mu_0}{L.v^*} + (V_G - V_{tsat})\theta_2} \quad (129)$$

where $u = 1 + \theta_2(V_G - V_{tsat})^2$.

4.1.1 Process flow and design of experiment

For this work, 2 lots have been studied. They all carry different process variations. Process flow of 28 nm FD-SOI technology is detailed in Figure 4-1.

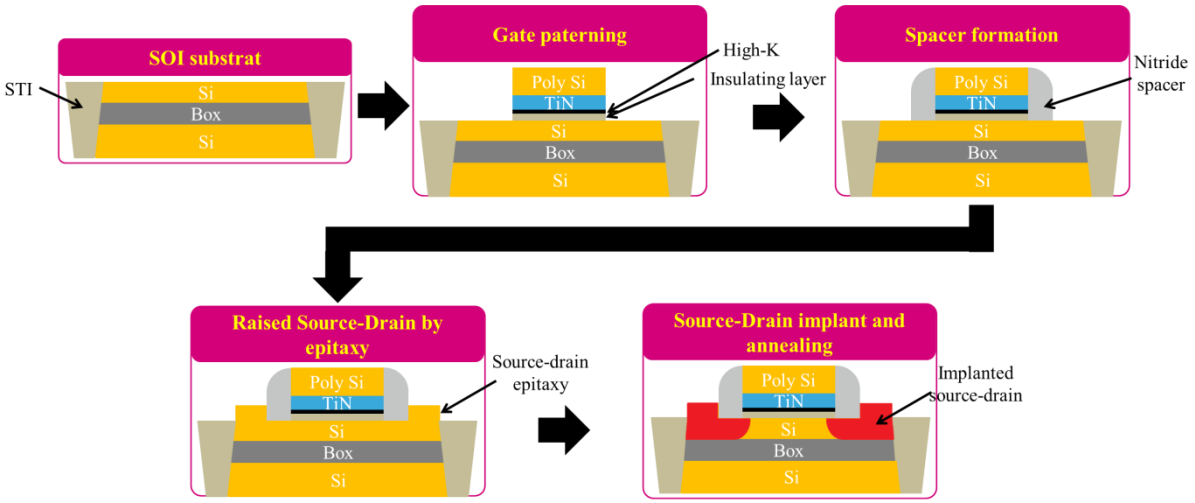


Figure 4-1: Schematic process flow of tested devices

Experiments focus on source-drain implant and anneal temperature. DOE summary is given in Figure 4-1:

Lot	A	B
Process	Low S-D implant	POR (800°C)

variations	POR	DSA 840°C
	High S-D implant	DSA 880°C

Table 4-1: Lot name related with process parameters modulation from POR

In Figure 4-1 POR stands for Process Of Reference. There is one POR for each lot. PORs are not identical since they belong to different lots that have not been treated at the same time. Thus, a POR is to be compared with process variation within the same lot. S-D implant stands for source-drain implant dose and energy. It refers to the last step in Figure 4-1 schematic when the dopants are implanted in the source and drain region. DSA stands for Dynamic Surface Anneal and is a rapid laser annealing treatment done after dopant implant **[144]**. Its aim is to activate implanted dopants without diffusing them. POR is exempted of this step.

Every process variation has been tested on several wafers (between one and three) and each wafer has been probed at least on 17 dies. Each site embeds many devices with different gate lengths. Each device drain current is probed a different gate biases. Gate length and probing biases of each device are summarized in Table 4-2 and Table 4-3.

Technology	28 nm FD-SOI	14 nm FD-SOI
Available gate Lengths [µm]	0.028	0.02
	0.030	0.024
	0.034	0.03
	0.12	0.06
	0.3	0.1
	1	0.3
		1

Table 4-2: Device gate length for which data are available.

For each of these gate lengths, drain currents have been measured in linear and saturation regimes at different gate voltages. These gate voltages are gathered in Table 4-3.

Technology	28 nm FD-SOI		14 nm FD-SOI	
Drain bias	0.05 V	1 V	0.05 V	1 V
Gate voltages for which data are available [V]	0.7	0.7	Vt_{lin}+0.3	0.4
	1	1	Vt_{lin}+0.5	0.8
	1.1	1.1	0.8	
			Vt_{lin}+0.7	

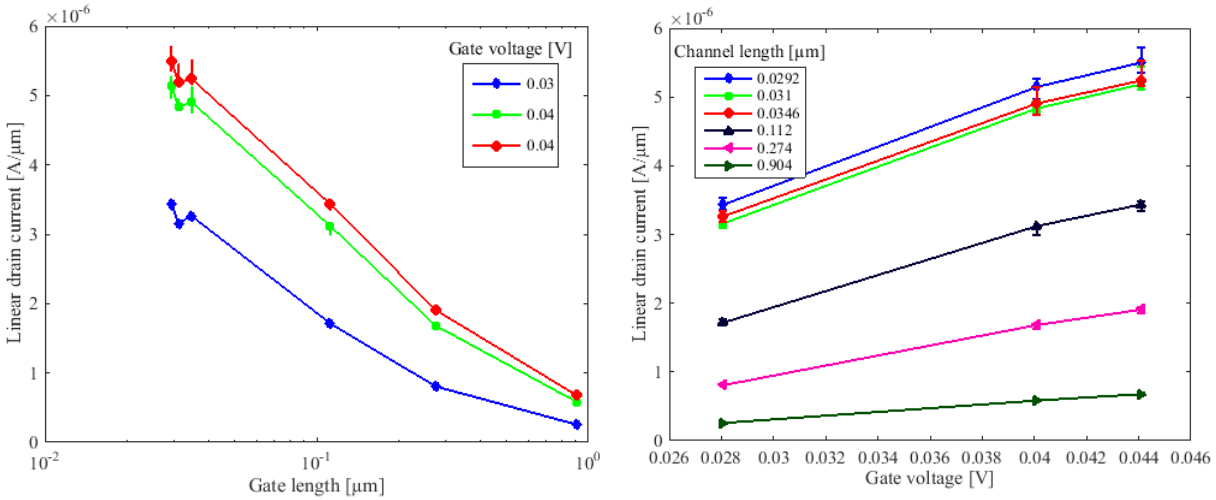
Table 4-3: Device gate voltages for which data are available.

Extraction is performed site by site. Thus we obtain for each wafer a distribution of model parameter. This dispersion gives an idea about the model parameters uncertainty at wafer scale.

4.1.2 Inference on process parameters effects on performance variations

4.1.2.1 Impact of source-drain implant energy and dose

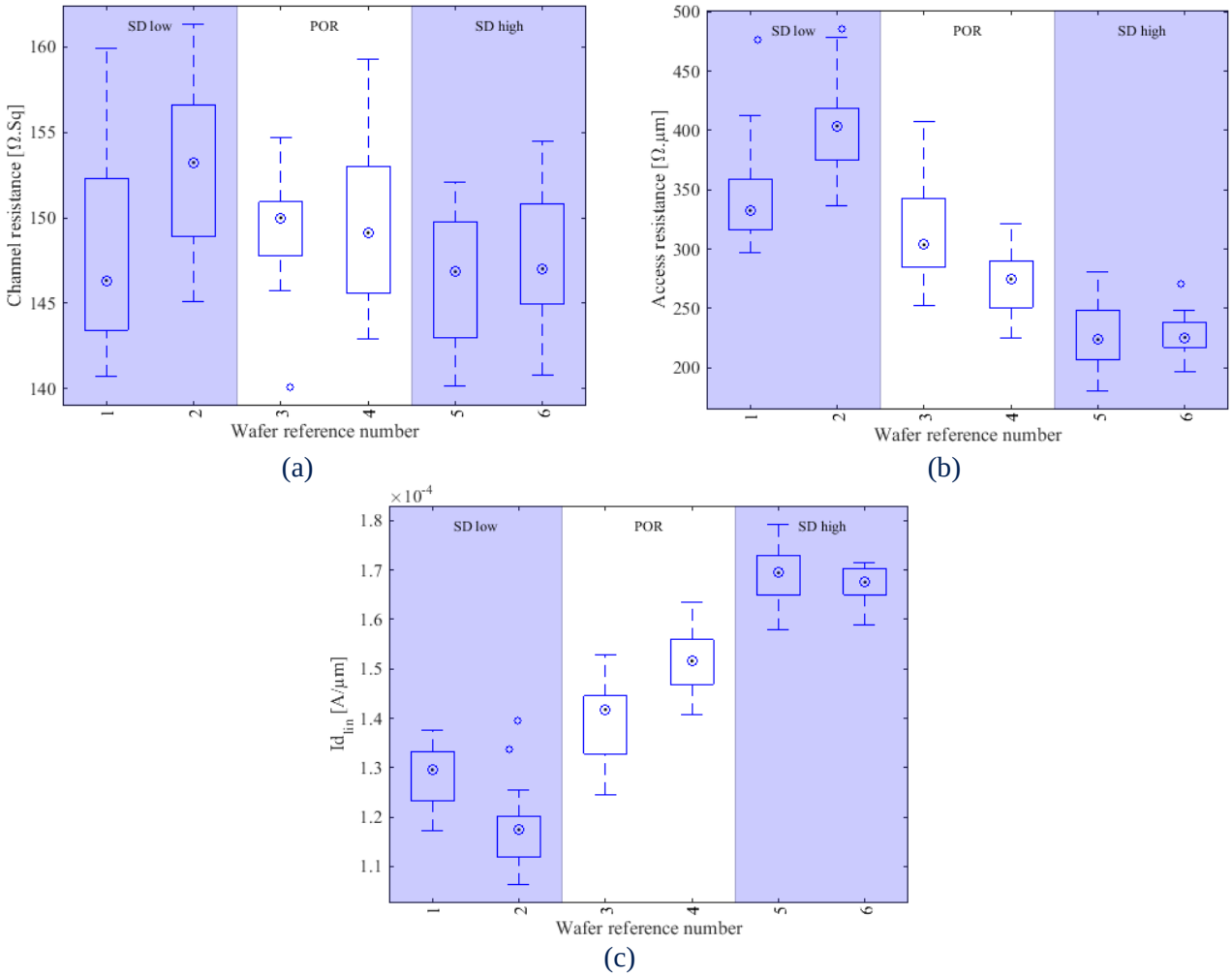
Here we investigate the effect of source-drain implant energy and dose variations, focusing on the results yield by lot A for nMOS devices. Changing the dose and energy of source-drain implant should only influence access region. Indeed, since transistor are built with a gate first process, implant only reaches the source and drain region. We might see a shift in the threshold voltage and in the carrier mobility if the dopants penetrate the metal gate as shown in **[145]**. These hypotheses will be discussed along with the extraction results.



(a) (b)

Figure 4-2: nMOS linear drain current against gate length (a) and gate voltage (b).

First, in order to assess the extraction robustness, Figure 4-2 shows the linear drain current model error (error bars) as a function of gate voltage (a) and channel length (b). For this figure, extraction has been performed on each site. Each point represents the median drain current value observed over the whole POR wafer and each error bar represents the standard deviation of the model error. A good adequacy is found.



(a) (b) (c)

Figure 4-3: Distribution of channel resistance (a), access resistance (b) and short channel device linear drain current at $V_G = V_{dd}$ (c) for each wafer of lot A.

Figure 4-3 shows the box plot of access resistance, channel resistance and linear drain current for each wafer of lot A. Channel resistance is the length dependent part of the total resistance whereas access resistance is what remains. This splitting of the total resistance is illustrated in (130)

$$R_{lin}.W = \boxed{R_0 + \frac{\sigma}{V_{GS}-V_{tLDR}}} + \boxed{\frac{L}{\mu_0 C_{ox}}\left(\frac{1}{\left(V_G-V_t-\frac{V_{DS}}{2}\right)} + \theta_2\left(V_G - V_t - \frac{V_{DS}}{2}\right)\right)} \quad (130)$$

Access resistance Channel resistance

Isolated dots are outliers. Data is spot as outlier if its distance to the median is larger than $q_3 + 1.5 \cdot (q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles, respectively. Figure 4-3 (c) shows a clear impact of process variation on linear drain current. This process variation affects only access and not channel resistance as shown in Figure 4-3 (a) and (b) as expected.

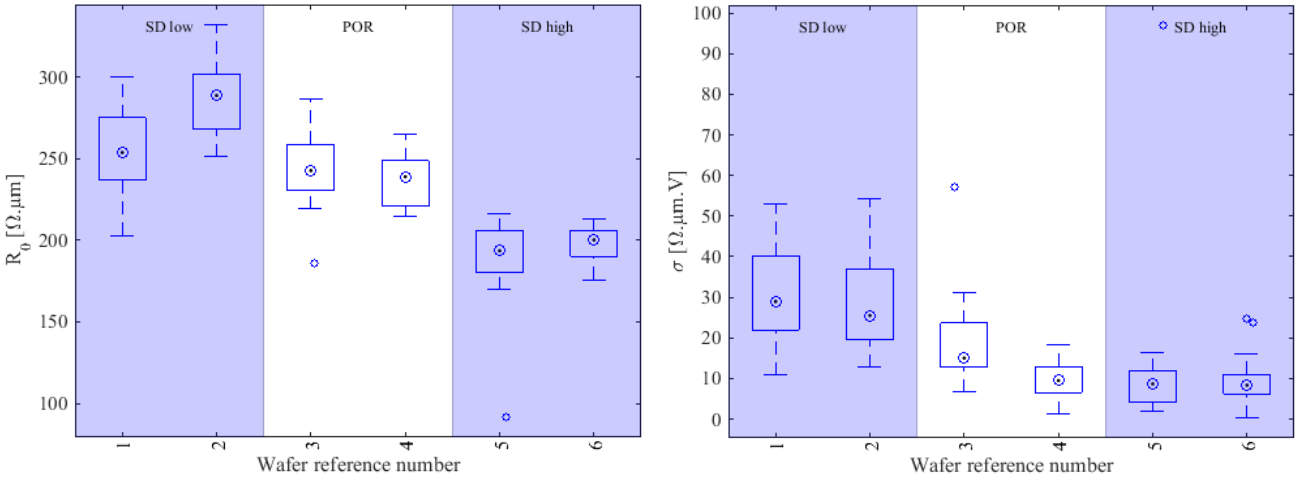


Figure 4-4: R_0 (a) and σ (b) distribution for each wafer of lot A.

In order to go deeper in the understanding of the impact of this process parameters, Figure 4-4 shows the distribution of R_0 and σ for each wafer. As expected, R_0 is clearly impacted by source-drain implantation and varies from 270 $\Omega.\mu m$ for lightly doped source-drain, down to 200 $\Omega.\mu m$ for heavily doped source-drain region. σ is impacted as well (from about 30 $\Omega.\mu m.V$ down to 10 $\Omega.\mu m.V$).

Figure 4-5 shows distribution of $\mu_0.C_{ox}$, θ_2, long and short channel Vt_{lin} for each wafer. As suggested by Figure 4-3 (a), channel model parameters are not affected by the process variation. Especially, Figure 4-5 (c) shows that long channel Vt_{lin} does not vary with process varaitions. Thus, dopants don't go through the gate to reach the channel.

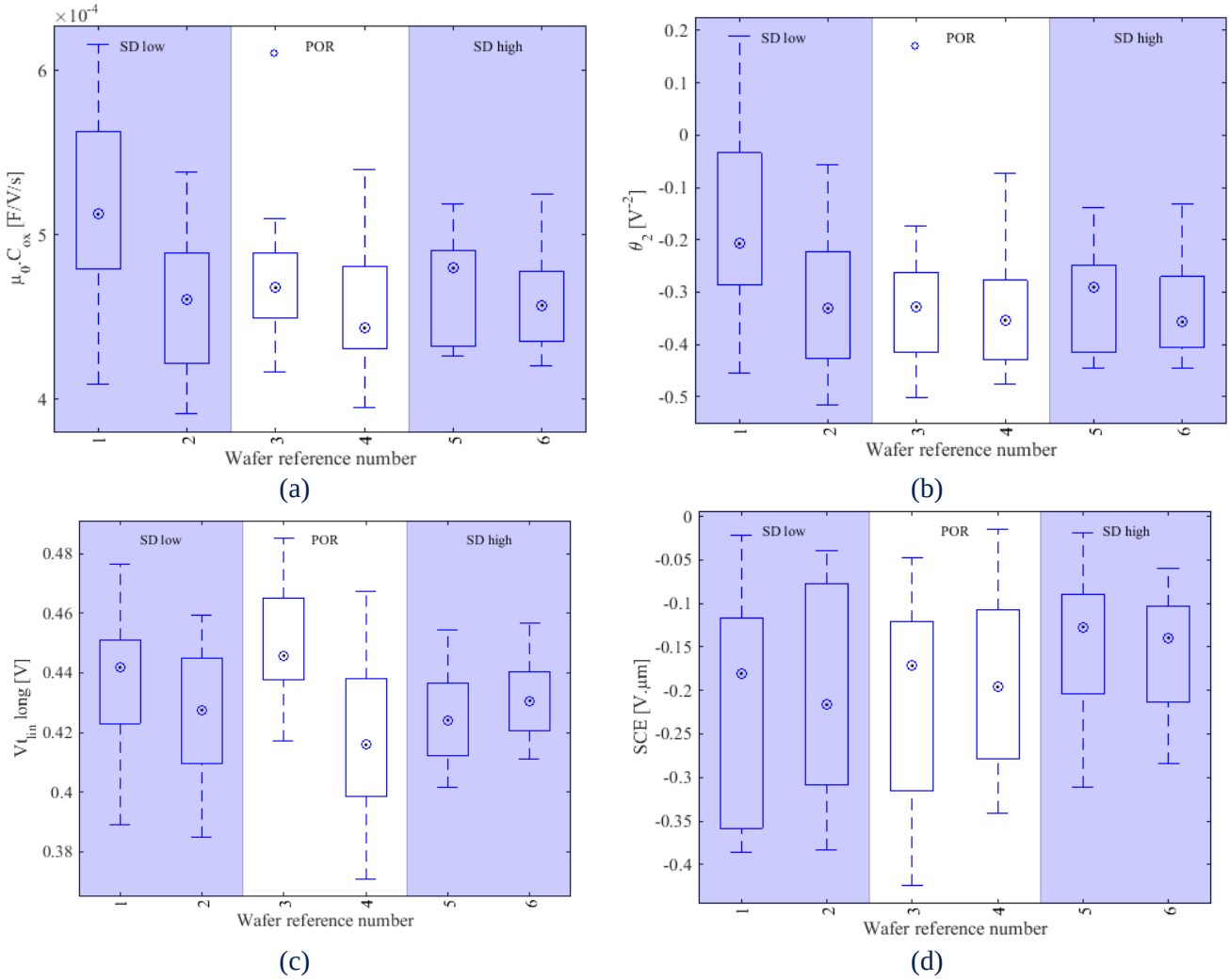


Figure 4-5: $\mu_0.C_{ox}$ (a), θ_2 (b), long channel Vt_{lin} (c) and SCE (d) distribution for each wafer of lot A.

Vt_{LDR} distribution for each wafer is presented in Figure 4-6. We see that it slightly depends on process variation, indicating that the LDR doping concentration has changed. This is coherent with process variation, thus for this case Vt_{LDR} extraction seems working.

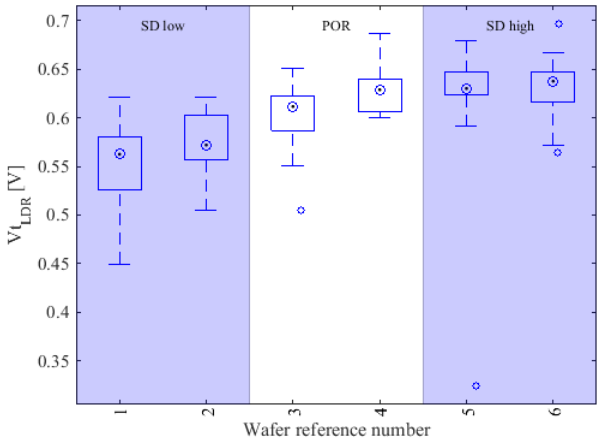


Figure 4-6: Vt_{LDR} distribution for each wafer of lot A.

Figure 4-7 shows the correlation between linear model parameters. The strong correlation between θ_2 and $\mu_0.C_{ox}$ implies that we cannot distinguish their contribution to electrical parameters variations. This correlation has been explained by the extraction range in §3.3.3. Since these parameters are not

implicated in above conclusions, the analysis still holds. On the contrary, other parameters are uncorrelated, confirming the robustness of the extraction and the reinforcing previous conclusions.

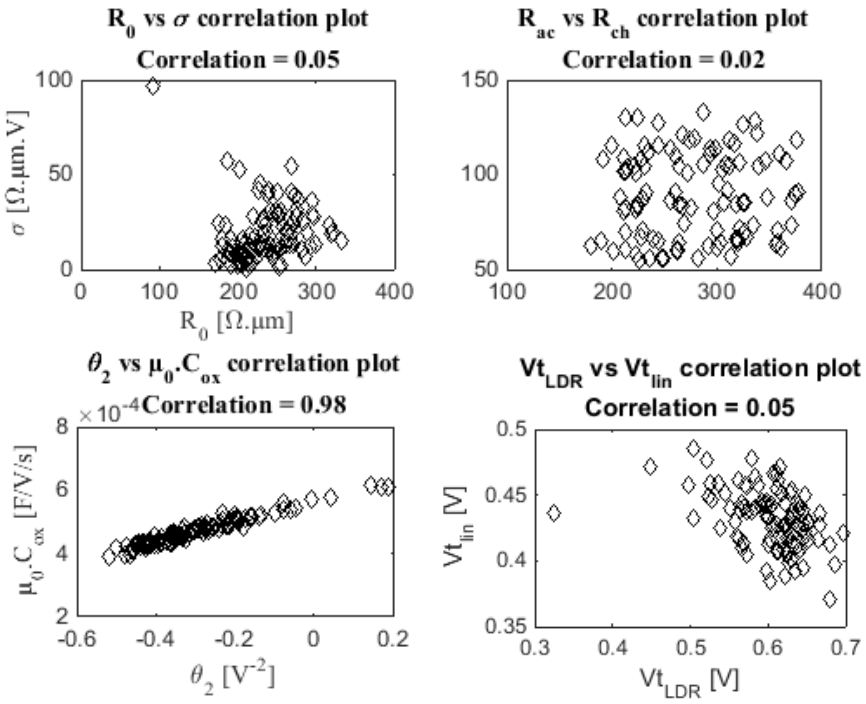


Figure 4-7: Linear model parameters correlation plot extracted over lot A.

4.1.2.2 Impact of DSA

In this section we investigate the effect of DSA on model parameters. DSA is aimed at activating dopants, avoiding migration. Thus R_0 should be lowered. LDR might be impacted by DSA through dopants activation but since no dopant migration is expected, this impact should be relatively low. Here we use the same equation to model the drain current than the one used previously. To verify the extraction robustness for this new lot, Figure 4-8 shows the correlation plot for model parameters.

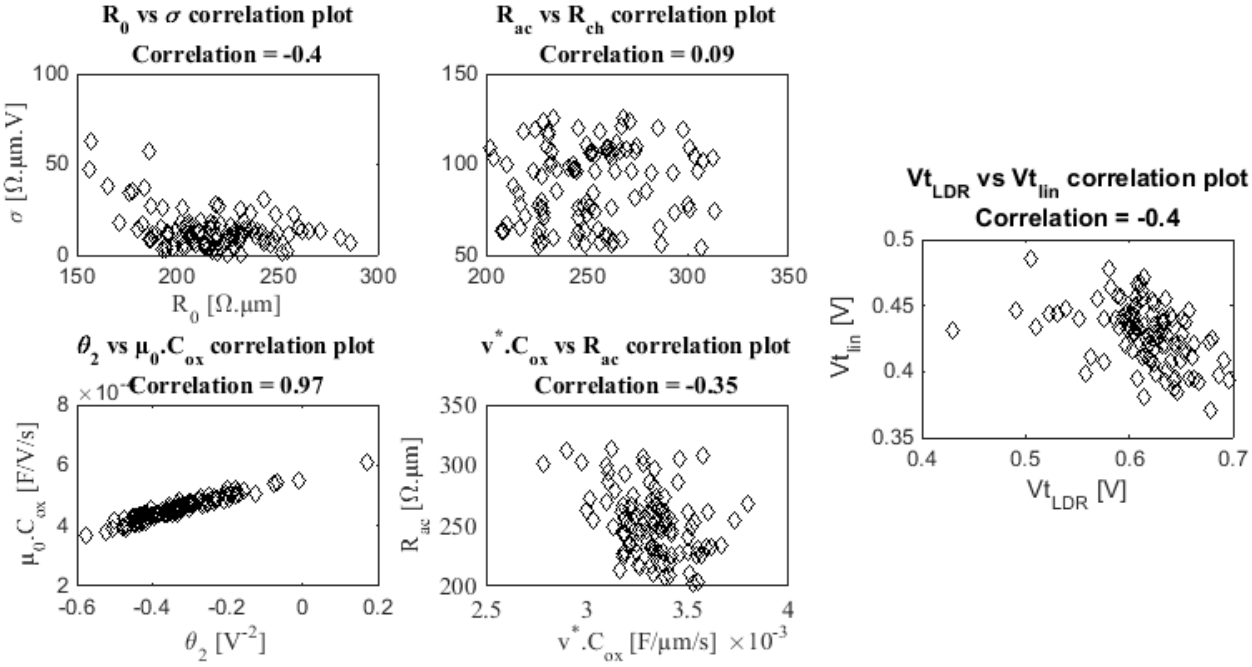


Figure 4-8: Model parameters correlation plot extracted over lot B.

Again we see that parameters are uncorrelated except for $\mu_0.C_{ox}$ and θ_2. Thus these parameters will not be distinguished in later analysis. Figure 4-9 shows I_{Dlin} variations against process variations. DSA tend to increase linear drain current.

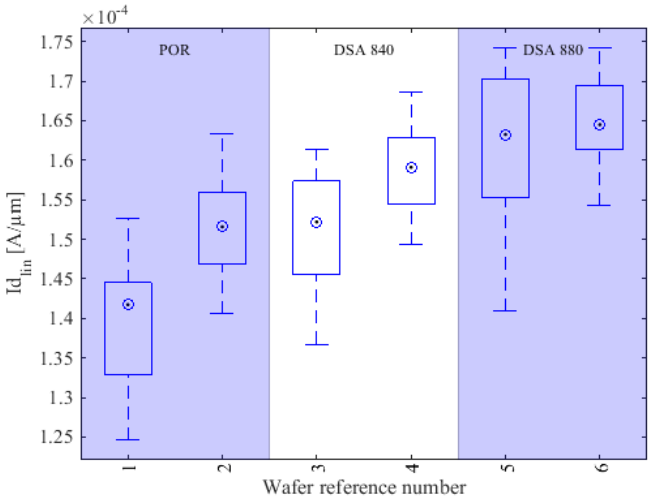


Figure 4-9: Short channel linear drain current distribution for each wafer of lot B.

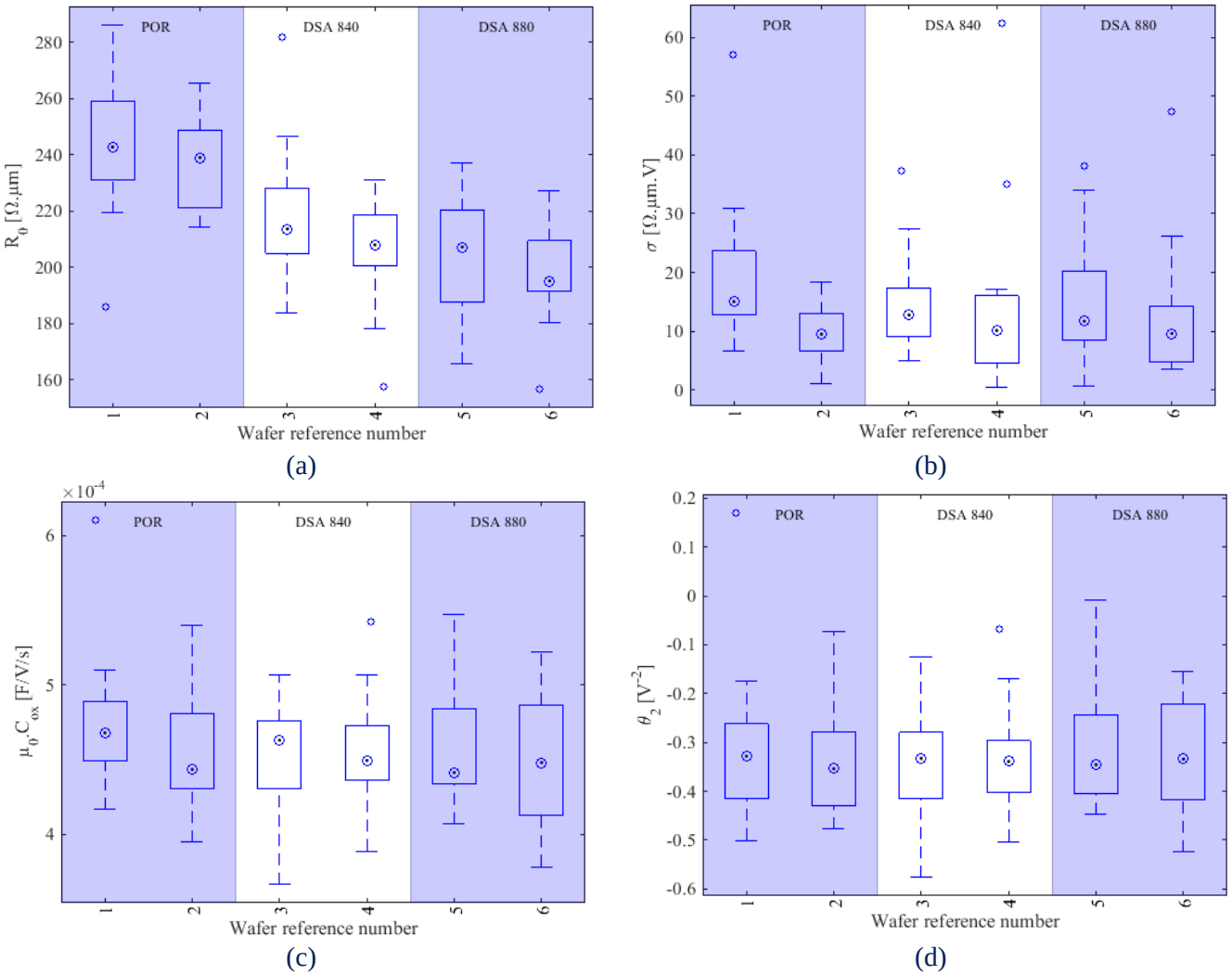


Figure 4-10: R_0 (a), σ (b), $\mu_0.C_{ox}$ (c) and θ_2 (d), dispersion for each wafer of lot B.

Figure 4-10 shows R_0, σ, $\mu_0.C_{ox}$ and θ_2 variation against process variation. R_0 is significantly reduced thanks to DSA meaning that all implanted dopant have not been activated during regular anneal steps. DSA helps activating them significantly. σ is not sensitive to DSA meaning that the junction did not move as expected and LDR dopants are already activated. Channel parameters are not sensitive to DSA either.

Figure 4-11 shows Vt_{LDR} distribution for each wafer. Vt_{LDR} is steady confirming the hypothesis that the junction has not moved. Since DSA does not induce dopant migration, it is expected. This is to be compared with Figure 4-6 where Vt_{LDR} changed due to higher implant does and energy.

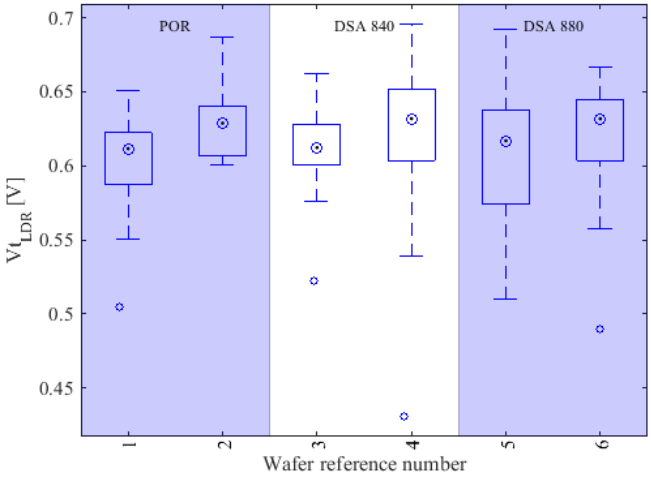


Figure 4-11: Vt_{LDR} distribution for each wafer of lot B.

Figure 4-12 shows the results of I_{Dsat} and $v^*.C_{ox}$ extraction for each wafer. I_{Dsat} is only slightly impacted by DSA. $v^*.C_{ox}$ seems to not be impact by DSA and there is no physical reason for v^* or C_{ox} to depend on the DSA. R_0 has a limited impact on I_{Dsat} but since it strongly depends on the DSA, it can explain the small I_{Dsat} dependence on DSA.

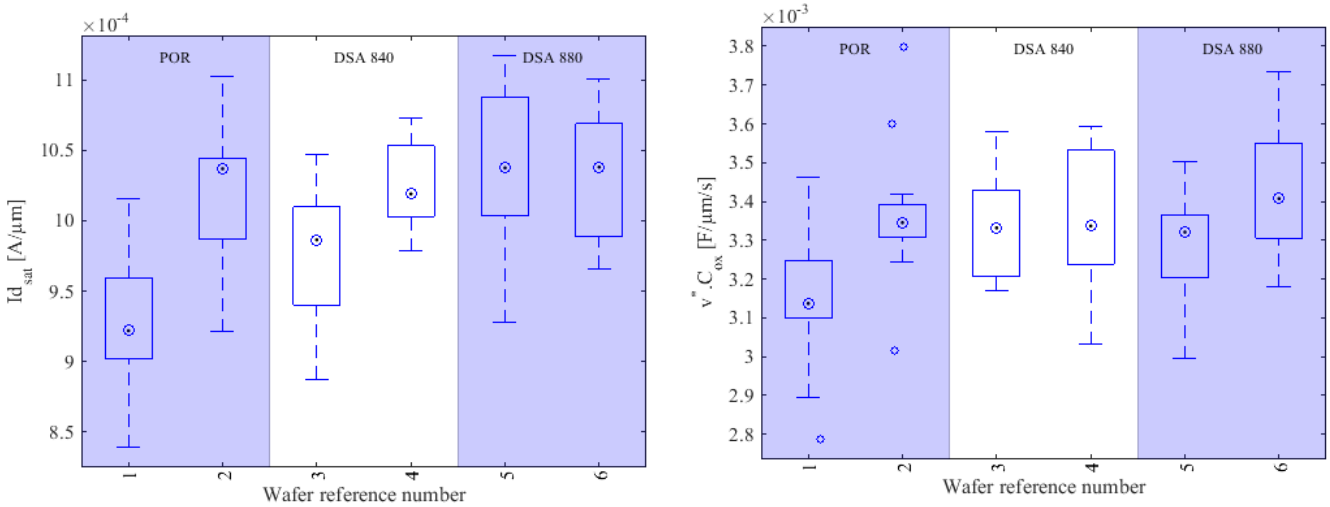


Figure 4-12: Short channel saturation drain current and $v^*.C_{ox}$ distribution for each wafer.

4.1.3 Conclusions about extraction on 28 nm FD-SOI devices measurements

In paragraph 4.1.2, we have applied the extraction procedure using 28 nm FD-SOI silicon devices measurements including variation of source-drain implant dose and energy as well as DSA. We have seen that extractions yield physically coherent results. R_0 is lowered by higher dose and energy

implant and by DSA. Both of these process parameters directly influence the active dopant concentration. σ only depends on dose and energy implant. If the dose is increased, the LDR is more doped. Moreover higher energy might have moved down the pic concentration of implanted dopants in the source drain region. This pic becomes then closer to the channel and dopants diffuse farther under the spacer and gate considering the same anneal treatment. Vt_{LDR} extraction has evidenced that the junction position is sensitive to implant energy and dose. On the contrary, DSA does not change the junction position and LDR doping concentration. This has been evidenced showing constant σ and Vt_{LDR} no matter if a DSA has been applied or not. Finaly, we can notice in both cases that $Vt_{LDR}>Vt_{lin}$. This suggests that LDR requires higher gate voltage to be inverted as discussed in §2.5.1. Thus the transistor could be underlapped.

4.2 Application to 14 nm FD-SOI technology

In this section, the extraction method is applied on 14 nm FD-SOI technology. For this work, a 16-wafers lot has been investigated. Process Of Reference (POR) wafer has been probed on 68 sites and the others are probed on 17 sites. First the model accuracy is assessed with correlation plots and model error evaluation in §4.2.1. The DOE along with the experimental setup are detailed in §4.2.2 along with the results and their interpretation.

4.2.1 Extraction accuracy assessment

In order to validate the capability of the extraction method to properly extract model parameters independently of each other, Figure 4-13 shows correlation plot of model parameters as well as access and channel resistance. In this figure, only the most strongly correlated couple of parameters are shown. Low correlation coefficients are found, emphasizing the robustness of the approach.

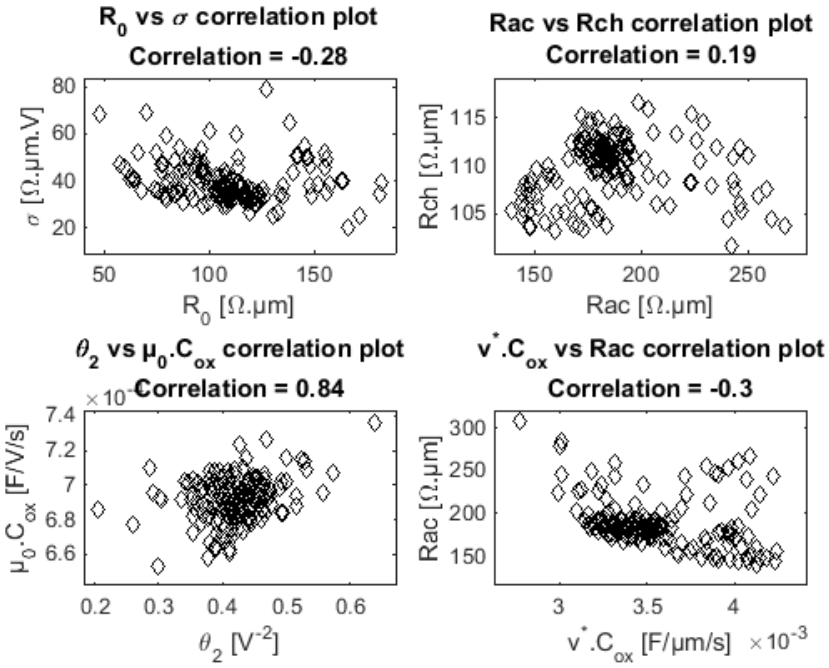


Figure 4-13: Correlation plot between channel and access resistance on POR wafer.

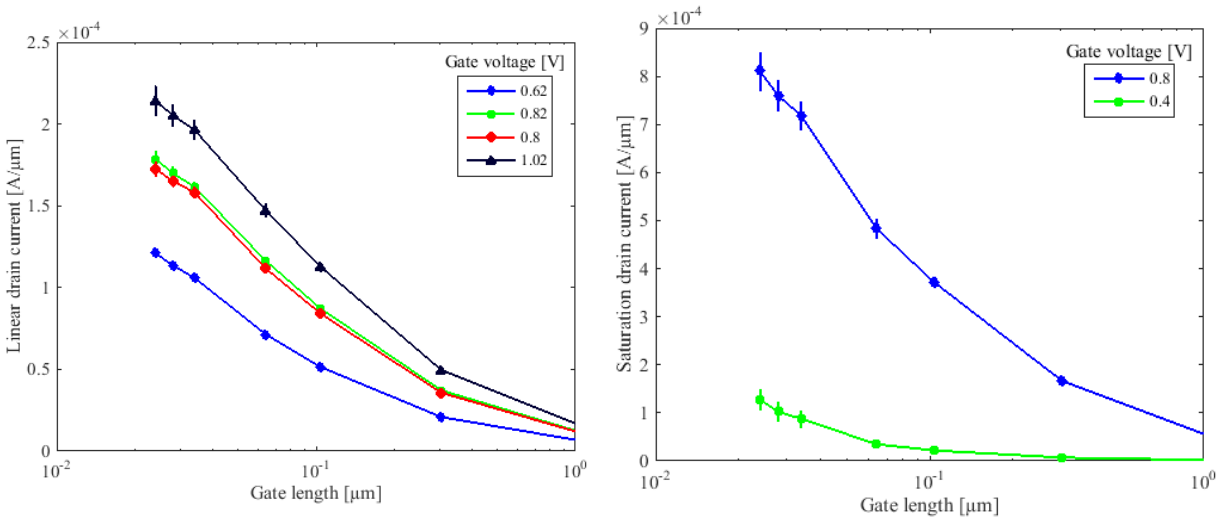


Figure 4-14: nMOS linear (a) and saturation (b) drain current against gate length. Symbols and lines represent measurements averaged over the POR wafer and error bars represent 3.$\boldsymbol{\sigma}$ of model error.

Figure 4-14 (a) shows the linear drain current model accuracy over the POR wafer for nMOS devices. Measured drain current is plotted against gate length for all V_G. Error bars show the standard deviation of the model error calculated over the 68 dies. In the same way as Figure 4-14 (a), Figure 4-14 (b) shows the saturation drain current model accuracy over the POR wafer for nMOS devices. Error is small confirming the ability of the model to predict measurements.

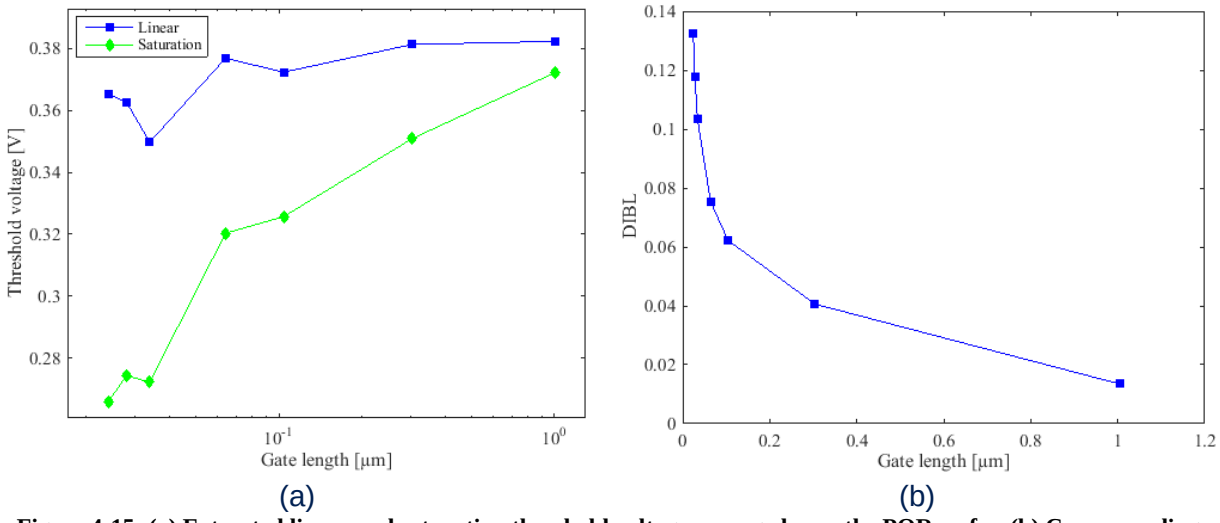


Figure 4-15: (a) Extracted linear and saturation threshold voltage averaged over the POR wafer. (b) Corresponding DIBL against gate length.

Figure 4-15 (a) shows the threshold voltage against the gate length, averaged over the POR wafer. A rough approximation of drain induced barrier lowering has been calculated using the following equation.

$$DIBL = \frac{Vt_{lin} - Vt_{sat}}{V_{dd} - Vd_{lin}} \tag{131}$$

Values for DIBL against gate length are shown in Figure 4-15 (b). These values are close to the one found in literature for FD-SOI technologies **[146]**.

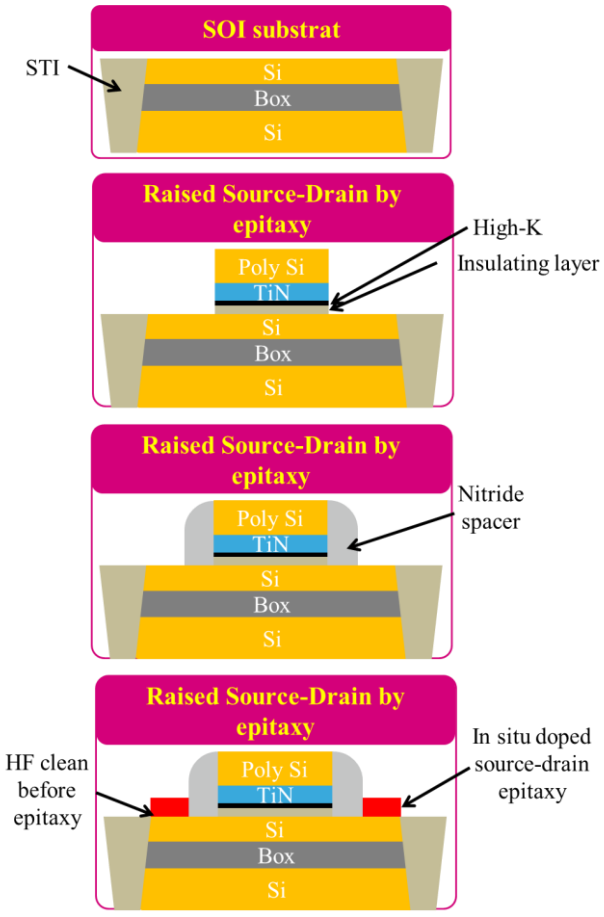
Table 4-4 regroups the average model parameters extracted over the POR wafer. v^* value is close to v_{sat} value found in literature **[147]**. A good mobility value is found as well.

Parameters	Values
R_0 [Ω. µm]	103
σ [Ω. µm. V]	37.4
$\mu_0[cm^2/V/s]$	199
θ_2 [V^{-2}]	0.46
v^* [cm/s]	$9.92 \cdot 10^6$

Table 4-4: Extracted model parameters averaged over the POR wafer.

4.2.2 Process flow and design of experiment

Process parameters have been varied from wafer to wafer. Figure 4-16 (a) shows the common process flow used to build devices under test. Process variations are localized at the end of the flow, during HF treatment and source drain epitaxial step. Modulated process parameters are the HF clean before source-drain epitaxy, epitaxial thickness, carbon and phosphorous dose injected during the epitaxial growth. Figure 4-16 (b) relates the wafer number with the process parameters modulation.



(a)

1	POR (reference)
2	
3	HF-
4	HF treatment before source drain epitaxy is shortened
5	HF- -
	HF treatment before source drain epitaxy is more shortened
6	SICP C+
7	Carbon is added during source drain epitaxy
8	
9	SICP C&P+
10	Carbon and extra Ph dose is added during source drain epitaxy
11	SiCP C&Pmax
12	Carbon and maximum Ph dose is added during source drain epitaxy
13	
14	SICP thick
15	source drain epitaxy is thicker
16	SiCP thin
	source drain epitaxy is thinner

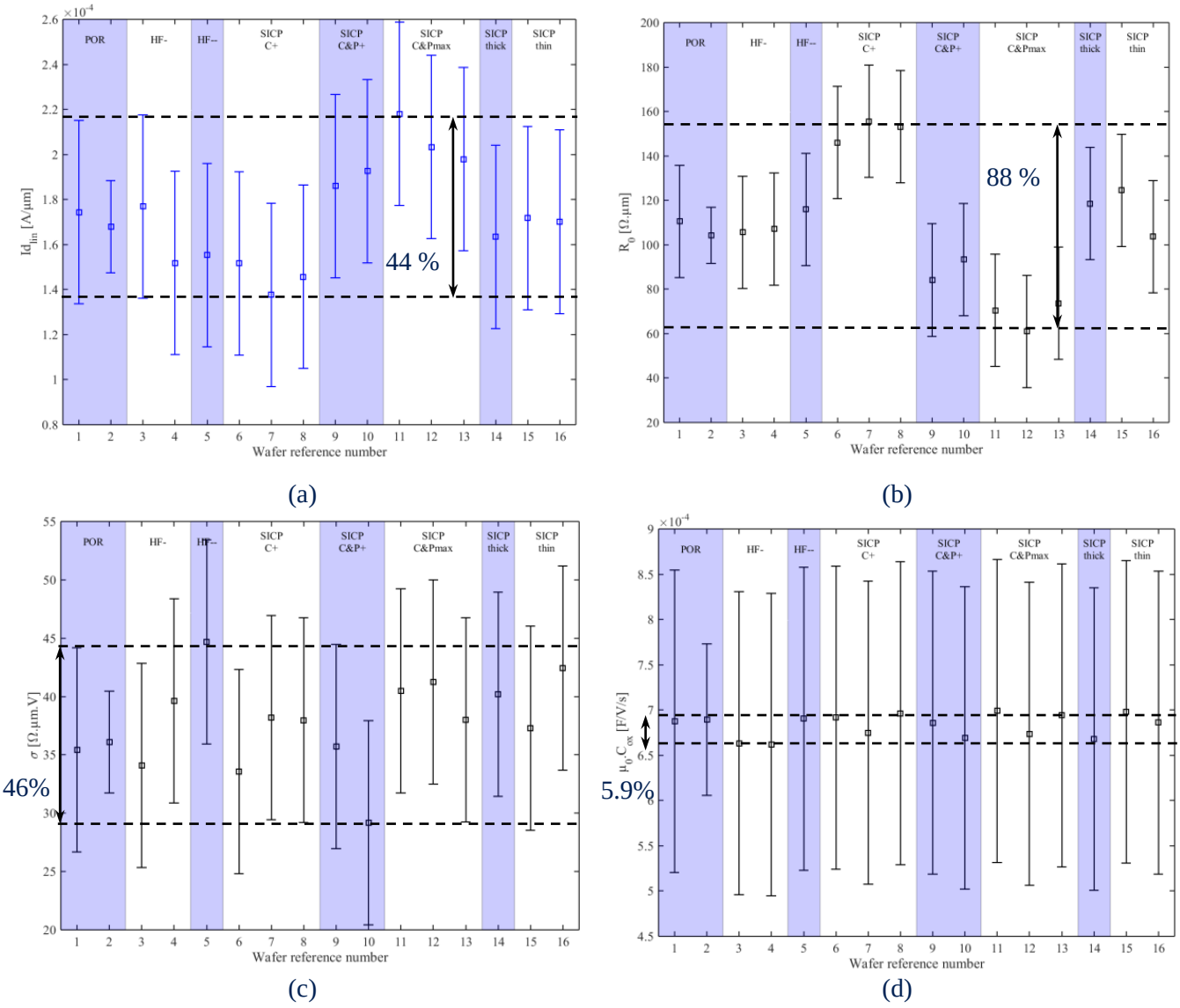
(b)

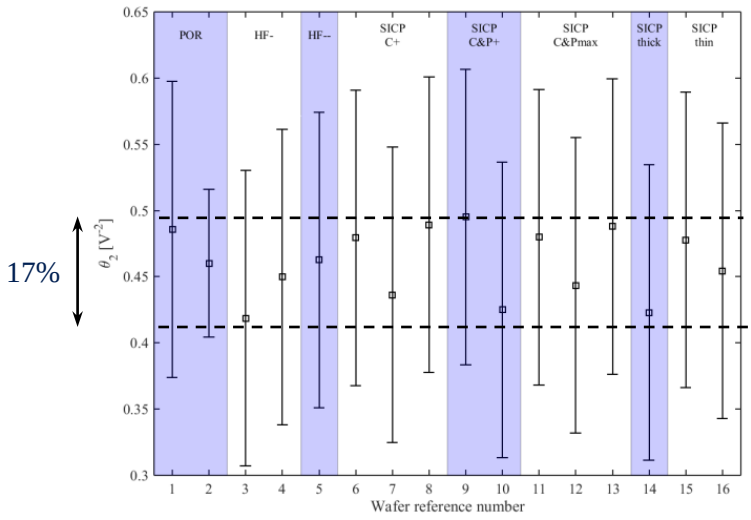
Figure 4-16: (a) Schematic process flow of tested devices [148]. (b) Wafer number related with process parameters modulation from POR

4.2.3 Inference on process parameters effects on performance variations

The impact of some specific process parameters on device performance is investigated here. To do so we use the extraction method introduced previously. Figure 4-17 (a) shows that drain current is rather sensitive to carbon and phosphorous dose. HF clean may induce a decrease in device performance but the effect is limited. Epitaxial thickness shows no major impact on drain current. Figure 4-17 (b) and

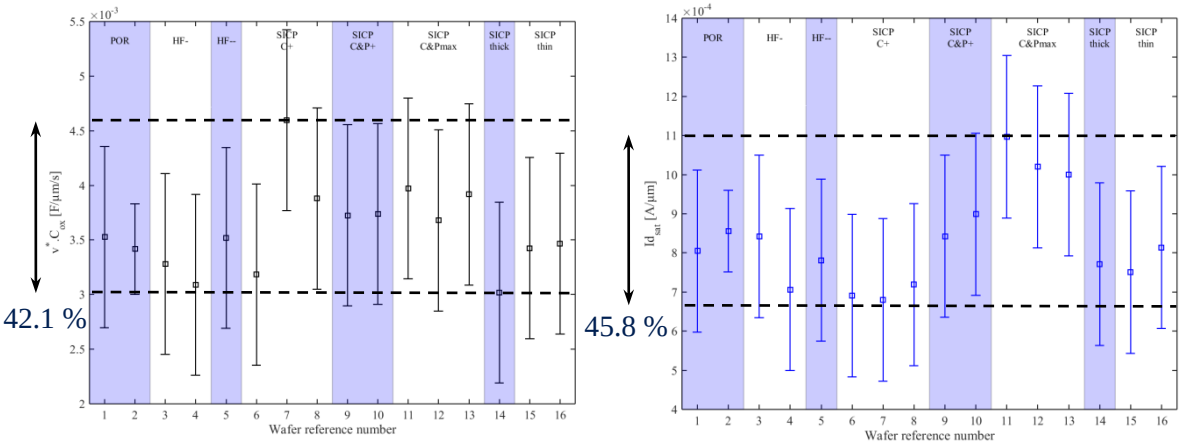
Figure 4-17 (c) focus on access resistance response with parameters R_0 and σ. Figure 4-17 (d) and Figure 4-17 (e) focus on channel resistance. We clearly see that R_0 is by far the first parameter that drives Id_{lin} wafer-to-wafer variation. Channel resistance is not much sensitive to process variations as expected since all process variations affect mainly the access region. Observed variations are in the range of wafer-to-wafer variability extraction accuracy. However the trend is clear for access resistance. Carbon raises R_0 since it slows down the dopant migration, whereas phosphorous dose reduces it by increasing the carrier concentration in the access region. σ seems only correlated to HF clean, indicating that it influences the junction position and the under spacer region. Short HF clean tends to degrade the contact quality between SOI and epitaxial raised source drain **[149]** creating silicon-oxide residues at the SOI/epitaxial interface. These residues act as defect sinks and can fix a large number of dopant and may induce cluster creation. Since these defects are fixed, they do not induce TED. Thus we expect dopants migration to be degraded. In our case, we see that shorter HF clean increases σ. TCAD extraction has shown that an increase of σ is due to a displacement of the junction away from the gate. Thus, it seems that shorter HF clean tends to move the junction away from the gate. This agrees with the hypothesis of enhanced cluster formation or sink for dopants, preventing them to diffuse toward the channel.





(e)

Figure 4-17: Drain current box plot (a), mean channel resistance (b) and mean access resistance (c) over each wafer for nMOS devices. Sq refers to an L/W normalization of the channel resistance. Only L=20nm shown.



(a) (b)

Figure 4-18: Wafer-to-wafer $v^*.C_{ox}$ (a) and I_{Dsat} (b) variations.

Figure 4-18 shows that $v^*.C_{ox}$ and I_{Dsat} variations are different depending on the process. In addition I_{Dsat} response to process variations is really close to I_{Dlin} response and v^*C_{ox} is not correlated to R_{sd} as shown in Figure 4-13. Thus $I_{Dsat,}$ as well as I_{Dlin} variations are mostly driven by R_0 variations.

4.2.4 Conclusion about 14 nm FD-SOI technology extraction

The model has been used to evaluate the impact of process variations on average MOS performance (R_{on}) over 16 wafers. A good fitting quality, as well as uncorrelated and physically relevant model parameter values, validates the model accuracy. Process variations considered are HF clean before epitaxial, carbon and phosphorous dose in source-drain region and epitaxial height. We have seen that considered process variations mainly affect R_0 and σ parameters (i.e. access resistance). Poor HF clean tends to act as a dopant sink, preventing them from migrating toward the channel. Thus it tends to make underlapped transistors and raises σ parameter. Carbon raises R_0 since it slows down the dopant migration, whereas phosphorous dose reduces it by increasing the carrier concentration in the access region.

4.3 Within-wafer variability modeling

In this section, we attempt to understand the relation between electrical and model parameters within-wafer variability. In order to address the within-wafer variability challenge, we will investigate three approaches: Monte Carlo draws, Forward and Backward Propagation of Variance (MC, FPV and BPV respectively). The last two methods have been widely investigated by McAndrew et al. **[150]-[151]** on BJT devices and MOSFET using PSP SPICE model **[152]**. We here apply them to model the transistor total resistance standard deviation on the POR wafer based on previously introduced device model. Results yield by different methods will then be compared.

4.3.1 Definition

4.3.1.1 Monte Carlo

In previous paragraph, we have introduced the analytical model and calibrated it on silicon based on the results yield by parameter extraction. Knowing the variability of model parameters, Monte Carlo method predicts the variability of electrical parameters (i.e. drain currents) by successively drawing normally distributed random sets of model parameters and computing electrical parameters using the analytical model. If the normality assumption is verified and the model parameters statistics is sufficiently accurate, electrical parameters statistics obtained by Monte Carlo should match silicon data.

4.3.1.2 Forward propagation of variance

Using Monte Carlo, the standard error of electrical parameters statistics is inversely proportional to the square root of the number of experiments. In other words, the larger the number of experiments, the more accurate the results is. This can lead to time consuming calculations. On the contrary, FPV formalizes Monte Carlo approach, giving the mathematical expression of electrical parameters statistics, knowing the model parameters statistics. We recall the basic equation of variance propagation here. Let's first call e_j the linear drain current (with j going over the 7 different channel lengths measured at 4 different gate voltages) and m_i the model parameters (with i going from 1 to 11 accounting for parameters R_0, σ, $\mu_0.C_{ox}$, θ_2 and the 7 Vt_{lin} of each channel length). Equation (132) relates model parameters covariance matrix σ_m^2 to electrical parameters covariance matrix σ_e^2:

$$\sigma_e^2 = J^T \cdot \sigma_m^2 \cdot J \tag{132}$$

where J is the sensitivity matrix of e with respect to m:

$$J_{i,j} = \frac{de_j}{dm_i} \tag{133}$$

This method propagates model parameters standard deviation using a first order expansion of the model. Thus model parameters variations should be small enough so that the model can be linearly approximated around the model parameters average. In addition, this limitation can be overcome by using second order sensitivity matrix. This approach has been investigated by McAndrew et al. **[153][154]**, however we will see that, in our case, a first order approximation is sufficient by comparing it with Monte Carlo which do not suffer from this drawback.

4.3.1.3 Backward propagation of variance

The two previous methods assumed that model parameters statistics are known with sufficient accuracy. Indeed, it can be accessed using extraction procedure introduced previously, based on full wafer electrical measurements. However, extracted statistics can be biased by the imperfection of extraction procedure. BPV is an alternative solution to estimate the statistic distributions of model parameters based on electrical parameters statistics without relying on extraction procedure. BPV provides σ_m^2 by mean of least square fit following (134):

$$\sigma_m^2 = (J \cdot J^T)^{-1} \cdot J \cdot \sigma_e^2(Vg, L) \cdot J^T \cdot (J \cdot J^T)^{-1} \quad (134)$$

This equation is simply the inverse function of (132), thus BPV require the same assumptions than FPV. This calculation is straight forward but depending on the size and condition number of J, the results can be numerically unstable. Thus, the results will be checked against Monte Carlo that does not suffer from such a problem.

4.3.2 Results of Monte Carlo vs FPV vs BPV vs silicon

4.3.2.1 Linear regime

In this section we compare the results yield by previously introduced methods in linear regime. Results are regrouped in Figure 4-19, where measurements are shown in blue. First, Monte Carlo method is applied using 10^5 draws of model parameter set. Random draws are based on model parameters statistics extracted using nonlinear extraction method (red symbols), using BPV (green symbols) and using linear least square fit (avoiding the second step of the extraction procedure: dark symbols). Statistics includes cross correlation between model parameters. Then FPV is applied with parameters statistics obtained using nonlinear extraction (red line), and using BPV (green line). Error bars represents the standard error about R_{lin} standard deviation.

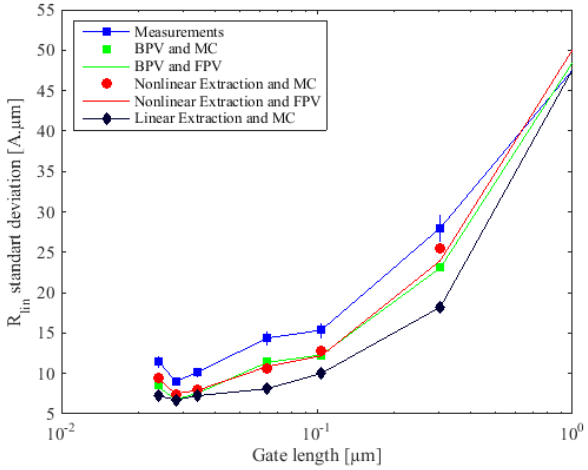


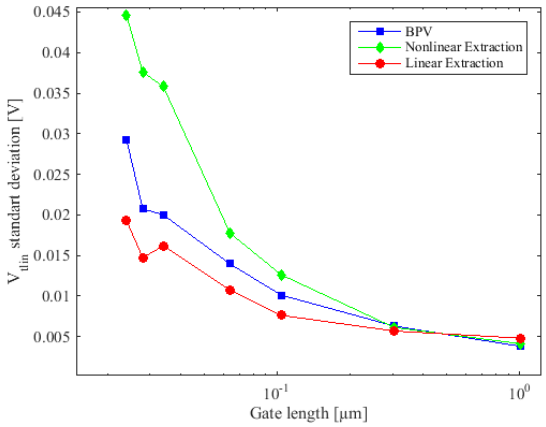
Figure 4-19: R_{lin} nMOS standard deviation over the POR wafer against gate length. Results include Monte Carlo, FPV based on parameters statistics extracted with nonlinear and linear least square fit and BPV.

First we see that nonlinear least square fit improves the results significantly compared to linear least square fit. It illustrates the bias that a poor extraction method can induce. Then Monte Carlo yields the same results than FPV. This means that model parameters dispersion is small enough to enable the first order approximation of the model done by FPV and BPV and the computational complexity is well handled. Figure 4-20 shows discrepancies between extracted model parameters using nonlinear and linear least square fit and BPV. Results don't match and BPV tend to yield larger model parameter

variability than direct extraction except for Vt_{lin}. Thus these three approaches are different even though BPV and nonlinear regression give close results for R_{lin} standard deviation.

	Extracted standard deviation	BPV results	Discrepancy (%)
R_0 $[\Omega.\mu m]$	9.86	20.4	69.7
σ $[\Omega.\mu m.V]$	6 89	14.3	70.2
$\mu_0.C_{ox}$ [F/V/s]	$7.80 \cdot 10^{-6}$	$7.55 \cdot 10^{-5}$	-3.27
θ_2 $[V^{-2}]$	$3.19 \cdot 10^{-2}$	$3.68 \cdot 10^{-2}$	14.2
$v^*.C_{ox}$ $[F/\mu m/s]$	1.61	1.61	0.1

(a)



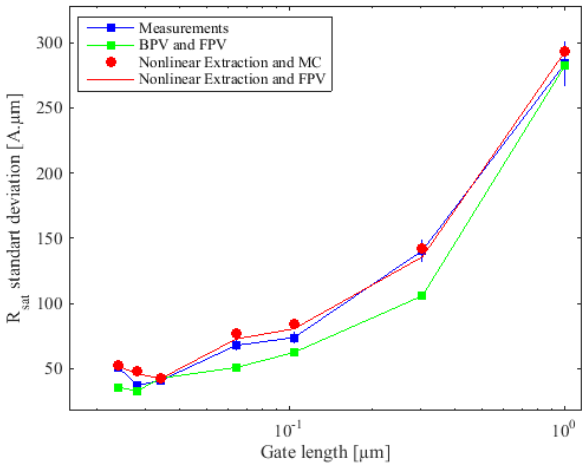
(b)

Figure 4-20: Model parameters standard deviation extracted using nonlinear optimization method and BPV.

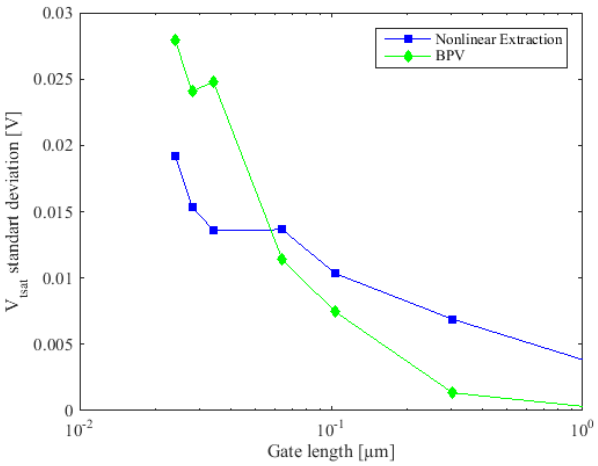
Comparing modeled and measured R_{lin} variability shows that the model makes systematical underestimation. Indeed the whole variability is not taken into account with this method. First the channel length is assumed to not suffer from any variability source. Second, using extraction procedure, within-die variability is not accounted for because all model parameters except V_t are unique and fixed for every devices of each site. The second point is not relevant for BPV because it does not rely on extraction procedure.

4.3.2.2 Saturation regime

Previous method has been applied here with measured saturation drain current to model R_{on} variability. In this case, BPV has been applied using R_{lin} and R_{on} measurements in $\sigma_e^2(Vg, L)$ in order to compute σ_m^2 that includes $v^*.C_{ox}$. The system is poorly conditioned compared with the case of linear regime. Indeed, singular value decomposition had to be applied to solve (134) because $J \cdot J^T$ becomes singular.



(a)



(b)

Figure 4-21: R_{on} (a) and Vt_{sat} (b) standard deviation depending on channel length using Monte Carlo and FPV and BPV against measurements.

R_{on} variability model results are shown in Figure 4-21 (a) where error bars represents the standard error of measured R_{on} standard deviation. We see here that the fit between model and measurements is

much better, suggesting a reduced impact of local variability on the saturation resistance. BPV yields poorer results compared to MC. This discrepancy arises from the calculation complexity.

Vt_{sat} standard deviation is plotted against gate length in Figure 4-21 (b). We see that again BPV method is not equivalent to extraction method.

4.3.3 Addressing channel length and local variability

In this section we investigate the influence of channel length and local variability on the different methods using synthetic data. To do so, three dataset of synthetic R_{lin} are randomly generated based on model parameters statistics found using nonlinear extraction over the POR wafer. The first set considers neither channel length nor local variability. The second considers channel length die-to-die variability ($\sigma L = 10\ nm$) and the last consider both intra-die variability for all model parameters except V_{tlin} and die-to-die channel length variability. For intra die variability, device to device model parameters have been modulated by 100% of the within-wafer variability.

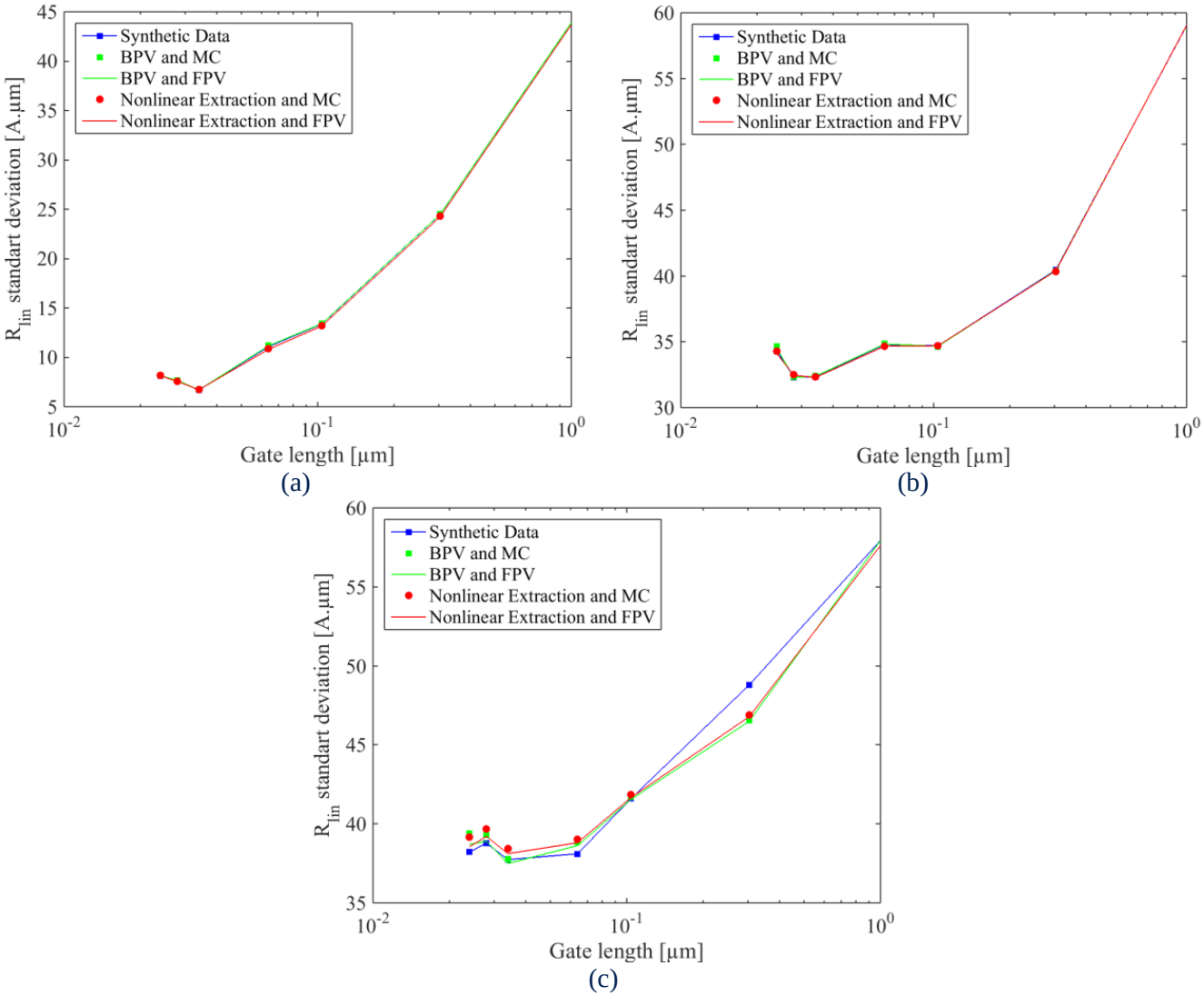


Figure 4-22: Synthetic R_{lin} variability against channel length. Synthetic data are generated with, (a) neither local nor gate length variability, (b) die-to-die gate length variability and (c) local and gate length variability.

Figure 4-22 shows synthetics R_{lin} variability against L along with the predicted variability using MC, FPV and BPV methods as in Figure 4-19. Figure 4-23 shows V_{tlin} variability used as input to the synthetic data generation along with the one predicted by extraction and the one predicted by BPV

method. Plots have been done based on the three synthetic data sets. We see that if there is no local variability, every method works fine and model parameters are well extracted. However when gate length and local variability is introduced, methods fail to track R_{lin} and Vt_{lin} variability properly.

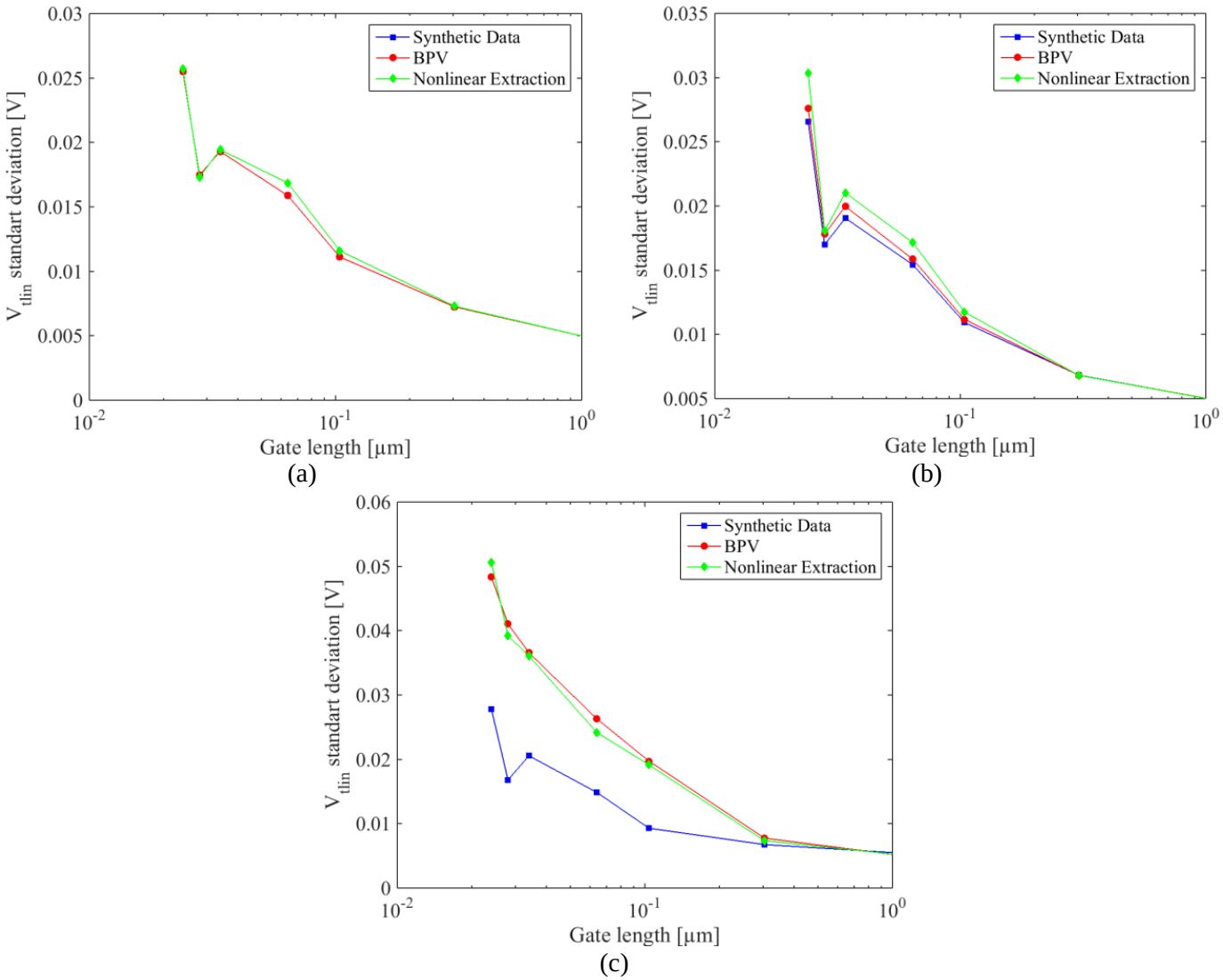


Figure 4-23: Synthetic Vt_{lin} variability against channel length. Synthetic data are generated with, (a) neither local nor gate length variability, (b) die-to-die gate length variability and (c) local and gate length variability.

Figure 4-24 shows the error on extracted model parameters using the three set of synthetic data. We see that without variability, model parameters are perfectly extracted as expected. Gate length variability only biases extraction of R_0, σ and Vt_{lin}. Local variability biases extraction of all model parameters. Thus local and gate length variability along with the simplicity of the model can explain the small discrepancies observed between measurements and model in Figure 4-19 and Figure 4-21 (a).

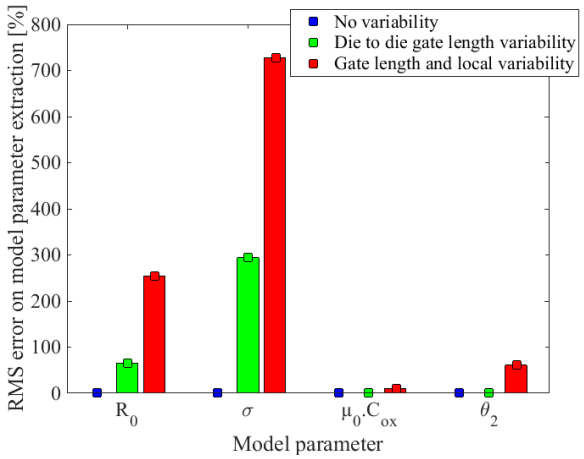


Figure 4-24: Root mean square error on extracted model parameters using synthetic data. Error is plotted in percentage of the input model parameters.

4.4 Conclusion

Following the introduction of model parameter extraction procedure in chapter 3, we have applied it on silicon measurements. 28 nm FD-SOI and 14 nm FD-SOI technologies have been investigated. It has been shown that model parameters variations depending on process variations are coherent and have been physically interpreted. A clear quantification of the impact of process variations has been enabled, showing that the method is efficient and robust while requiring only few measurements, making it suitable for industrial application.

Studying 28 nm FD-SOI using model parameter extraction enabled quantifying the impact of source drain implant dose and energy as well as DSA step. We have seen that extractions yield physically coherent results. Highly doped source-drain region resistance R_0 is lowered by higher implant dose and energy and by DSA. Both of these process parameters directly influence the active dopant concentration. This means that highly doped source-drain region has remaining inactivated dopant before DSA. DSA activates them successfully. On the contrary LDR resistivity represented by σ is only dependent on implant dose and energy. Indeed DSA does not induce dopant migration and thus doesn't move the junction further toward the channel. Moreover this means that LDR dopants are already well activated before DSA and DSA has no activation effect. However Vt_{LDR} extraction has evidenced that the junction position is sensitive to implant energy and dose. $\mu_0.C_{ox}$, θ_2 and Vt_{lin} have been shown to be constant, meaning that dopant does not penetrate into the metal gate or channel. All these sensitivities can be quantified easily using this technique, bringing valuable information in terms of device optimization.

Studying 14 nm FD-SOI technology, it has been possible to evaluated the impact of HF cleaning time before epitaxy, carbon and phosphorous dose during in situ doped raised source-drain epitaxial as well as epitaxial thickness. Carbon has shown to increase R_0 by reducing dopant migration whereas increased phosphorous dose decreases R_0 by raising the active dopant in the highly doped source drain region. Poor HF clean tends to act as a dopant sink, preventing them from migrating toward the channel. Thus it tends to make underlapped transistors and raises σ parameter

In a second step, within-wafer variability has been investigated on 14 nm FD-SOI technology. Monte Carlo, forward and backward propagation of variance have been conducted in order to model this variability. It has been shown that linear model linear drain current variability is slightly underestimated. BPV and direct extraction showed close results in term of linear drain current variability however corresponding model parameters variability yield different results. It has thus been

suggested that local variability and channel length variability are responsible for these discrepancies (that are not properly taken into account using direct extraction or BPV). This interpretation has been reinforced by the fact Monte Carlo draws used to forward propagate the model parameters variability extracted using BPV and direct extraction gives the same results than FPV. This leads to infer that the discrepancy does not come from a violation of normality and linear local approximation hypothesis. In order to verify that channel length and local variability are responsible for observed discrepancies between measurements and model, their impact on the model has been assed using synthetic data and showing that it induces errors and can thus explain it.

Chapter 5 :
Process compact model

In the previous chapter we have successfully built the required tools to map the relation between model parameters and electrical performances. We have shown how to draw all the benefits of these tools. In particular it has been possible to get valuable insights into the device characteristics and understanding of process variations impact on the device functionalities. One last step is required in order to complete the model construction that relates electrical and process parameters. Indeed we miss the link between model parameters and process parameters (see the first stage of the PCM diagram in the introduction chapter of this manuscript Figure 1-3). The aim of this chapter is to show how to build this kind of model that is called Process Compact Model (PCM) in our context. Three methods to build PCM will be tackled: i) stepwise regression, ii) LASSO and iii) LARS. PCM construction is a thorny task and should be carried out cautiously. A misuse of these methods can lead to strongly biased and unreliable PCM. Indeed, in order to ensure a proper output, each of these methods requires to be calibrated before use. In order to calibrate and test the model efficiency, we will use cross validation tests. We have kept three variants of this kind of test: k-fold Cross-Validation (k-fold CV), Leave-One-Out Cross-Validation (LOOCV) and bootstrap. These methods give an estimate of the PCM accuracy and propensity to be predictive. Thus calibrating stepwise regression, LARS or LASSO consists in running these tests using different calibrations. The best calibration is the one that optimizes both accuracy and the propensity to be predictive. PCM construction process is depicted in Figure 5-1.

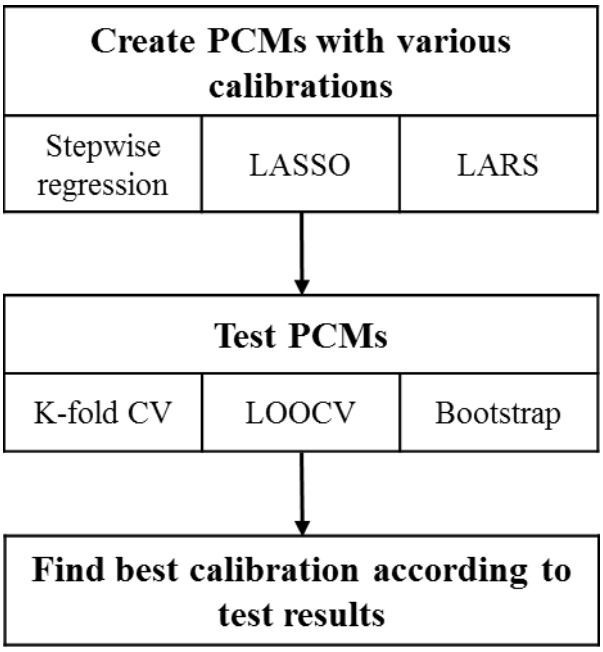


Figure 5-1: PCM construction process flow

PCM will first be defined in §5.1. We will see in which context this tool can be useful. Strategies to calibrate and test the robustness of PCM such as K-fold CV, LOOCV and bootstrap are introduced in §5.2. PCM construction methods like stepwise regression, LASSO and LARS will be detailed in §5.3. In §5.4 we will build PCM to link extracted model parameters using TCAD, introduced in previous chapter, with process parameters. In this paragraph we will show that using the PCM construction process flow (see Figure 5-1) is mandatory for silicon applications. Simpler approach would fail in this task because it implies dealing with ill-posed problem that requires variable selection and dealing with noise and variability. Using this PCM construction procedure, in §5.5, we will construct a PCM at the wafer scale and show that it can model efficiently within-wafer variability. This model will be used afterward in order to give guidelines to optimize within-wafer variability. We will see that wafer scale PCM can only account for process parameters that exhibit large dispersion at wafer scale. In order to

build PCM that includes larger process variations, the same procedure will be carried on a full DOE, in §5.6, with process parameters variations. The impact of local random variability, within wafer variability and measurement noise will be investigated using synthetic data. Recommendations will be given about the experimental setup required in order to build PCM with sufficient robustness and minimum error. A summary of this chapter is proposed in §5.7.

5.1 Process compact model (PCM) definition and context of use

5.1.1 Definition

Traditionally, definition of PCMs are models that relate process and device electrical parameters through a set of analytical functions, allowing manufacturing engineers to gain insights into device electrical parameters sensitivity to process variability in an extremely fast and robust manner **[155]**. This is agreement with the aim of this thesis. However, in the context of this chapter, "PCM" will also be used to designate the model that maps the relationships between process and model parameters. The simpler concept of PCM is illustrated in Figure 5-2:

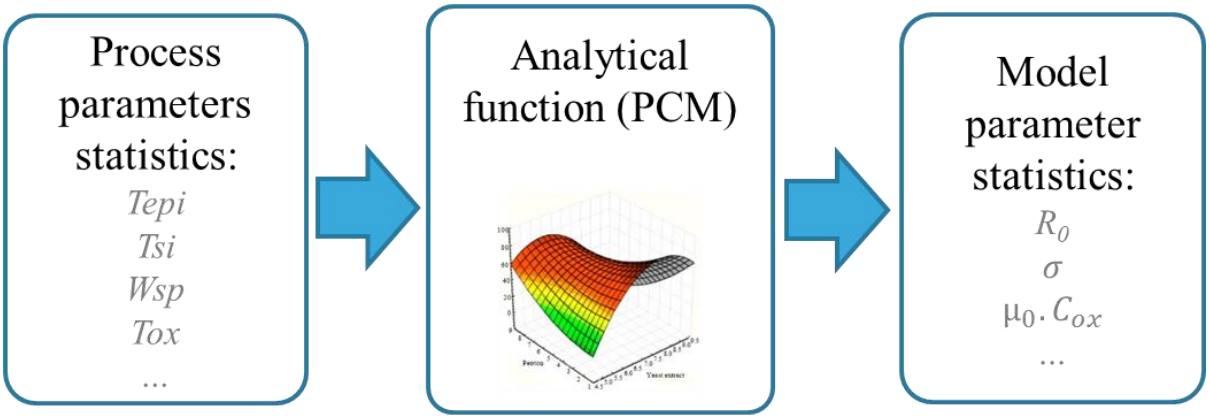


Figure 5-2: Illustration of the basic concept of PCM

As depicted, this PCM relates process and model parameters through analytical functions. This is not to be confused with compact model that relates electrical parameters with model parameters, where model parameters are not necessarily process parameters. Indeed compact models often relie on model parameters such as threshold voltage, DIBL, or subthreshold slopes that can have complex relationships with process. On the contrary, process parameters are geometrical or physical quantities that can be straightforwardly accessed through process adjustment (i.e. the epitaxial layer thickness Tepi can be modulated simply by varying the deposition time during the process). As well, PCM should not be confused with process and electrical simulation tool such as TCAD which use finite element algorithm to compute the physics and the electrical properties of the device. These simulations rely on numerical computation of complex physical models whereas PCM only use analytical fast computing models.

This kind of model (its construction and application) belongs to a mathematical field called data mining or statistical learning. In order to introduce properly the different mathematical tools used in this work, we provide some definitions here.

- Observations (noted $\hat{Y}$): They are what we want to model. Depending on the model considered, it can be either electrical measurements or simulations or quantities issued by extraction procedure (e.g. R_0, σ, $\mu_0 . C_{ox}$, V_t,…). These quantities are not known unless measured, simulated or extracted. In our case observations are extracted model parameters.

- Responses (noted Y): They are outputs of the model. They are the same quantities as observations and supposed to be as close as possible to them.
- Variables or predictors (noted P): They are inputs of the model (i.e. process parameters). They are known beforehand and set by the user.
- PCM coefficients (noted β): Fixed parameters used to calibrate the PCM in order to fit observations. These parameters are determined during the PCM building phase.

Following the mathematical formalism, parameters to be modeled called "observations" (noted $\hat{Y}$) and process parameters called "predictors" (noted P) are related via the PCM (noted f) that satisfies:

$$\hat{Y} = f(P, \beta) + e \tag{135}$$

And

$$Y = f(P, \beta) \tag{136}$$

where e is the residual between responses and observations. Residual can be due to model deficiency or noise induced by measurements $\hat{Y}$.

5.1.2 Applications and benefits

PCM have not been extensively used in literature. Thus it can be difficult to understand its benefits and applications. In order to illustrate this, two studies found in literature are developed here. They use different kind of PCM in order to model electrical parameters.

Considering the complexity of physical and electrical mechanisms underlying state-of-the-arts MOSFETs, simple analytical functions are not suited for PCM. Literature reports the use of Feed-Forward Neural Network (FFNN) to overcome this complexity and build suitable PCM for emerging devices **[156]**. This study was based on TCAD simulations and aimed at investigating process variability in nanowire FinFets. The technology being not mature enough, only TCAD simulations can provide a good prediction of electrical and process relationships for this kind of device. However statistical investigations require a large amount of experiments and TCAD simulations are too much time consuming to give timely answers. In this context PCM has been used since they can meet the expectations. Their PCM could predict full I_D-V_G's starting from model parameters such as channel length, gate length and oxide thickness.

In another context, Kakehi et al. **[157]** have applied PCM to model the relationship between gate length, halo dose, RTA spike and the threshold voltage. Although they did not explain precisely how the PCM is built, it might be simpler than FFNN since it only relates 4 parameters. Moreover V_t is simpler to model that full I_D-V_G. They used it in combination with feed-forward process control in order to reduce die-to-die, wafer-to-wafer and lot-to-lot V_t variability. In Feed-Forward Process Control **[158][159]**, process conditions at n+1 step are varied so that the impact of the n^{th} step variability is minimized. Determining process variation at n+1 step is done using PCM.

To conclude, PCM provides a powerful tool for statistical analysis and process optimization. These two cases introduced above showed that, depending on the context and which parameters are supposed to be related, the strategy to build the PCM changes. In the first case the PCM is able to predict the full I_D-V_G but is based on model parameters that cannot be straightforwardly accessed through process adjustment (like electrical channel length). In the second case, PCM's input parameters are truly process parameters but its capability is limited to V_t prediction. In our study we will see how we can combine the efficiency of our compact model with a user friendly PCM that relates process and model parameters to get the full mapping of the device, from process to electrical parameters.

5.2 Methods to evaluate the accuracy of a model

Different methods to build PCMs will be introduced further. These methods need to be calibrated and will not lead to the same results in most cases. Thus a method should be determined to select the best model. Efficiency of the model is evaluated using two criterions: the mean square error (MSE) between model and observations and the model variance (that is the variance on extracted model coefficients β). This paragraph introduces some basic methods to evaluate the model efficiency.

5.2.1 Validation test

Considering a set of observations $\hat{Y}$ composed of n elements, the most straight forward way to evaluate model error is to build the model using $\hat{Y}$ and then calculate the model error following:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{137}$$

where Y is the model response and n the number of observations. However, using this technique, the same observation set is used to build the model and to calculate the model error. Thus model error does not reflect the ability of the model to predict observations not used to build the model within the observation domain. Model efficiency estimation is thus strongly biased using only MSE as indicator. The alternative is to split the observation set into a training dataset and a validation dataset. The model is then built on the training dataset and the model error is calculated on the validation dataset, which is different from the training dataset. This approach is called validation test. The training dataset domain is similar to the test dataset domain in order to ensure a proper coherence of test with the model. The respective size of training and validation dataset is set by the user. This choice leads to a tradeoff. Indeed if the validation set is small then, the model is built upon almost all the observation set, minimizing the model variability but model error estimation is subject to a large uncertainty. On the other hand, if the validation test is large, the uncertainty about model error is minimized but model variance increases.

5.2.2 Cross-validation [160]

Cross validation is an alternative to validation test. It consists in splitting the observation dataset into a training and validation dataset many times to improve the estimation of model error and allow estimation of model variance.

There are two common way to proceed. The first method is called Leave-One-Out Cross-Validation (LOOCV) and the other is k-fold Cross-Validation (k-fold CV). These methods are explained in the following paragraphs.

5.2.2.1 Leave-One-Out Cross-Validation (LOOCV) [160]

Considering n observations, LOOCV method assigns all but one observation to the training dataset. Test dataset is composed of only one observable. This method ensures building the best model out of available observations but leads to high uncertainty on model error since it is calculated on only one observation. In order to get the most accurate model error estimate, the procedure is repeated n times, each time choosing a different observation for the test dataset. The LOOCV estimates for the mean square error is the average of these n error estimates, n being the number of observations:

$$LOOCV\ MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2 \quad (138)$$

Although the first aim of LOOCV is not to evaluate the model variance, it can be used to do so. Indeed, at each iteration, the training test is different thus the method yields n models. It is then possible to evaluate the model variance on this set of n models. To do so, the total model variance is calculated as the average of each coefficient variance as:

$$LOOCV\ Model\ Variance = \frac{1}{n}\sum_{i=1}^{n} Var(\beta_i) \quad (139)$$

where Var stands for variance. LOOCV MSE and LOOCV Model Variance are not used to build a model by estimating β values but they are used to evaluate the model accuracy and variance. The lower they are, the more accurate and less variable the model is.

5.2.2.2 K-fold Cross-Validation (k-fold CV) [160]

An alternative for LOOCV is k-fold CV. In this approach, the entire observation dataset is split into k subset of the same size. One of them, called Y_k, is taken as the test dataset and the other are regrouped to form the training dataset. As LOOCV method, this procedure is repeated k times, each time using a different subset for test dataset. The k-fold CV estimate of the mean square error is the average of these k error estimates:

$$\text{k-fold CV MSE} = \frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}}\left(\mathrm{Y_k} - \widehat{\mathrm{Y_k}}\right)^2 \quad (140)$$

If k=n then this method is perfectly identical to LOOCV, however choosing $k \ll n$ ensure less uncertainty in model error estimation. Meanwhile the model variance might increase significantly if k becomes very small. In practice k-fold CV is applied using k=5 or k=10.

Model variance can be estimated using k-fold CV as well, following the same approach than LOOCV. Here there are only k model generated thus, the model variance estimate is less accurate. The model variance formula yields:

$$\text{k-fold CV Model Variance} = \frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}}\mathrm{Var}(\beta_\mathrm{i}) \quad (141)$$

5.2.3 Bootstrapping **[161]**

Bootstrapping is a resampling method that aims at estimating properties of an estimator (e.g. mean, variance, confidence interval, standard error …) without assumption about the distribution. Considering a sample $\hat{Y}$ of n observations, the method consists in resampling many times $\hat{Y}$ by successively drawing n elements in $\hat{Y}$ with replacement. Bootstrap sample can have the same element more than once. Thus every bootstrap sample can be different. For each new sample generated, the estimator is calculated. This operation yields an empirical distribution for the estimator whose properties can be estimated. For example, population mean can be estimated as the average of the

bootstrap samples means, and standard error about population mean can be estimated as the standard deviation of the bootstrap samples means.

If the population has a normal distribution, then bootstrap is just a more complicated way to derive properties of the estimator distribution. However bootstrapping provides results that hold even if the distribution is unknown.

In the context of building a PCM, bootstrap method will provide estimation about the model variance. To do so let's consider a dataset of k predictors P and n observations $\hat{Y}$:

$$z_i = [\hat{Y}_i, P_{i,1} \dots P_{i,k}] \tag{142}$$

with i going from 1 up to n and k the number of different predictors. Model parameters β can be estimated building the PCM using z. Following bootstrap method, the observation $z_1, z_2, \dots, z_n$ can be resampled B times. Bootstrap samples are collected and named $z_{b1}^*, z_{b2}^*, \dots, z_{bn}^*$ with b going from 1 up to B. Model coefficients are then computed for each of the bootstrap samples, producing B set of bootstrap model parameters, named β_b. Using the model parameters distribution we can evaluate their standard error.

However, directly resampling z implicitly treats the model parameters P as random rather than fixed. We may want to treat it as fixed since we choose it beforehand and it is not subject to any uncertainty. To do so the method consists in:

- Estimating the model parameters and calculate the response Y and residual E_i for each observation.

$$E_i = \hat{Y}_i - Y_i \tag{143}$$

- Generating bootstrap samples of residual E_{b1}^*, E_{b2}^*,.., E_{bn}^*.
- Calculating the bootstrap observation Y_{bi}^* subtracting E_{b1}^* to $\hat{Y}_i$.
- Extracting the bootstrap model parameters β_b from Y_{bi}^* and P.

Standard errors on model parameters are then calculated from β_b distributions. This alternative is valid only if the functional form of the model is correct and if residues are identically distributed over P.

5.3 Methods to build PCM

In this paragraph, different approaches to build PCM are introduced. These methods consist in two steps, first selecting the relevant predictors that will enter the model, second the model is created by fitting observations with these predictors. The global approach is detailed in §5.3.1, then two categories of methods are investigated, first subset selection in §5.3.2 and then shrinkage method in §5.3.3. More recently developed approach using hybridization of the two previous methods are introduced in §5.3.4. These methods are systematically tested against synthetic data in order to investigate their different behavior, strengths and weaknesses.

5.3.1 Global approach

Building a PCM consists in finding a model that relates output parameters with process parameters. The most straightforward way to build empirical linear model based on observation is ordinary least square fit. This approach is adapted only if predictors are all relevant (i.e. they are actually correlated

with the observation) and not correlated between each other, the problem is not ill-posed (i.e. there are much more observations than predictors) and observations are not subject to significant amount of noise or uncertainty. However this is not our case since there are more than 300 process steps for the Front-End-Of-Line (and thus at least as much process parameters). Moreover not all of these steps have an impact on electrical properties. For example, different dopant implantations are used for nMOS and pMOS devices. When pMOS source-drain implant is processed, nMOS devices are protected. Thus considering the pMOS source-drain doping implantation energy or dose as an input process parameter for nMOS device PCM is a mistake. This kind of selection is based on expert knowledge and is the first selection that should be made. Then, if we consider only relevant process parameters (those that actually play a part in the considered device building process), only a few of them will have a significant impact on the output parameters. Considering that the amount of observation is very limited the problem is ill-posed and ordinary least square is not an appropriate method. Thus variable selection must be performed. There are different methods to achieve this kind of task. Among them we distinguish two main categories: subset selection and shrinkage method. These methods are introduced in the next paragraphs.

In order to test their ability to successfully find relevant predictor despite the presence of noise and correlated parameters in an ill-posed problem, we build a predictor matrix P composed of 50 randomly generated predictors and 25 observations. The polynomial formula that links P and Y is arbitrarily set to:

$$\begin{cases} Y = 3 + 1.5 \cdot P_1 - 2 \cdot P_2 + 3 \cdot P_3 - P_4 + P_5 \\ \mathrm{P}_6 \dots \mathrm{P}_{50} = \text{random predictors not used to compute Y} \end{cases} \quad (144)$$

Only 5 predictors are used to generate the observation. The others are fake predictors. Noise is then simulated by adding a normally distributed amount to the observation. The noise level (i.e. the standard deviation of noise signal) is set to 10% of the observation. All of these PCM construction methods will be used on synthetic observations in order to check if they are able to find back the polynomial formula. These methods rely on user defined parameters that must be calibrated. To do so we will use cross validation and bootstrap methods introduced previously. It will also be used to compare the efficiencies of each method.

5.3.2 Subset selection [162]

This approach consists in determining the minimum set of predictor that explains the observation variance. Following approaches will be introduced: best subset selection, forward and backward selection as well as a hybrid method that relies on both backward and forward selection.

5.3.2.1 Best subset selection [162]

Considering that our problem is to build the best PCM composed of p predictors, the most straightforward way to do so is to compute every possible model and compare their efficiency. The most efficient model will contain only relevant predictors. This technique is called best subset selection. If the investigation is limited to linear model only, then the number of possible model depending on p is given by the following formula:

$$N_{models} = \sum_{i=1}^{p} \binom{p}{i} \quad (145)$$

In this formula i is the number of parameters included in the model. Considering all cases, i ranges from 1 parameter (a constant for example) up to p parameters (the complete model that includes every parameters). Then, if the model has i parameters, there are $\binom{p}{i}$ possible combinations and as much possible model. Thus, for 10 predictors (including the constant) there are 1023 possible models. This is already large if the model is time consuming to compute and test. If we want to consider second order polynomial models, then the possible number of models increases drastically and yields:

$$N_{models} = \sum_{i=1}^{p+p^2} \binom{p+p^2}{i} \tag{146}$$

where $p + p^2$ is the sum of the linear (p) and quadratic (p^2) parameters. Again, if p=10, there are $1.3 \cdot 10^{33}$ possible models. The problem is too large to be treated in a reasonable amount of time. This is why this method will not be used in practice.

5.3.2.2 Forward stepwise regression [162]

In order to alleviate this problem, the stepwise regression has been proposed in the 60's. We distinguish forward and backward stepwise regression. Forward stepwise regression starts by the simplest possible model that is a constant. Then predictors are added one by one, each time selecting the one among the remaining predictors that best explains the residue (in other words that gives the greatest additional improvement to the fit). There are many ways to determine the best predictors to be included in the model. Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R-square, Akaike information criterion, Bayesian information criterion, Mallows's Cp, PRESS, or false discovery rate **[163]**. In our case we will use F-test.

In order to compute F-test, three quantities related to the model are required: the Sum of Square Error (SSE), the Sum of Square Regression (SSR) and the degree of freedom (df):

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{147}$$

$$SSR = \sum_{i=1}^{n} \left(Y_i - mean\left(\hat{Y}\right)\right)^2 \tag{148}$$

$$df = n - p \tag{149}$$

In (147) and (148), n is the number of observations and p the number of predictors. At the i^{th} step of the model construction, there are i predictors included in the model and p-i predictors not yet included. SEE, SSR and df are calculated for the model with i parameters and for the model with i+1 parameters. The model with i+1 parameters is the model with i parameters to which we added a selected predictor among the one not yet included in the model. These quantities are called SSE_1, SSR_1, SSE_2, SSR_2, df_1, df_2. Subscripted indices 1 and 2 stand for the model with i and i+1 parameters respectively. Then the F value is calculated as

$$F = \frac{SSR_2 - SSR_1}{SSE_2} df_2 \tag{150}$$

This F value is calculated p-i times, each time using a different predictor for the model with i+1 parameters. This is how a F value can be associated with each parameter not in the model. The predictor to be added for the next step is the one that yields the smallest F value.

A criterion is needed in order to determine whether the next predictor to be added is relevant or not. This is the stopping criterion. The simplest stopping criterion consists in comparing the smallest F value with a critical F value F_{crit}. F_{crit} is calculated as the inverse cumulative probability distribution of Fisher statistic for a probability set by the user. Usually the probability is taken between 0.85 and 0.95. if $F<F_{crit}$ the algorithm can continue otherwise it stops because the predictor does not significantly improve the model prediction. This probability threshold can be set using cross-validation method.

5.3.2.3 Backward stepwise regression [162]

As an alternative, F-test can be applied to decide which parameter to remove instead of being added. This is called backward stepwise regression. This method starts by creating the most complete model (that includes all the p predictors). Then for each predictor of the model, F-test is calculated considering the full model and the one where the predictor under consideration is removed. The larger F value designates the predictor to be removed.

In this case stopping criterion is reached if $F<F_{crit}$ (where F_{crit} is calculated with a probability chosen between 0.05 and 0.15). Alternatively it can be also checked using cross-validation.

5.3.2.4 Hybrid approach between forward and backward stepwise regression [162]

The main flaw of forward stepwise regression is that, adding a predictor to the model changes the F value of every predictor already in the model. Thus, at the n^{th} step, a predictor that has been previously added to the model can, at this step, yield an F value larger than F_{crit}. Even though this predictor used to be the one that best explained the observation variance (at the step $i < n$), it might not explain it anymore.

To overcome this kind of problem, based on forward stepwise regression, hybrid approach consists in checking the F value of every predictor (including the one already in the model) at each step and deciding whether to add or remove predictor depending on their respective F values. Different approaches have been proposed to formalize this decision process. In our case, the predictor which has the larger difference between its F value and F_{crit} is chosen to be either removed (if $F> F_{crit}$) or added (if $F<F_{crit}$).

Matlab implemented stepwise algorithm **[164][165]** has been used to perform stepwise regression on synthetic data, using the whole dataset as a training dataset. In order to illustrate the behavior of this method, Figure 5-3 shows the estimated predictor coefficient β depending on parameter F_{crit}.

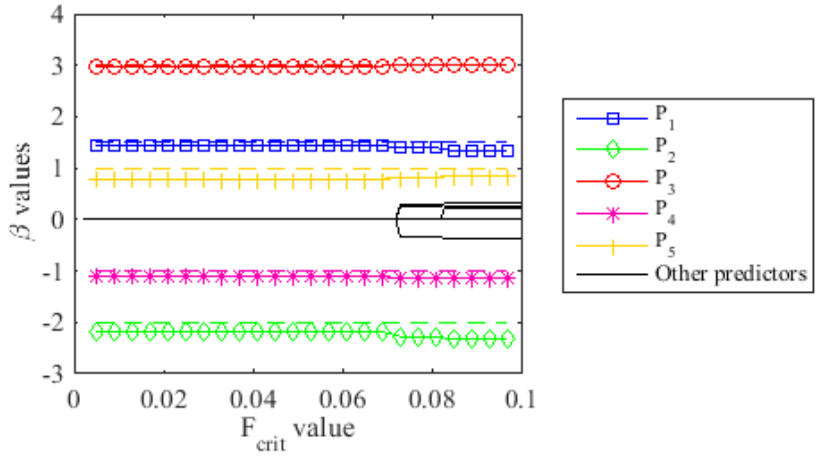
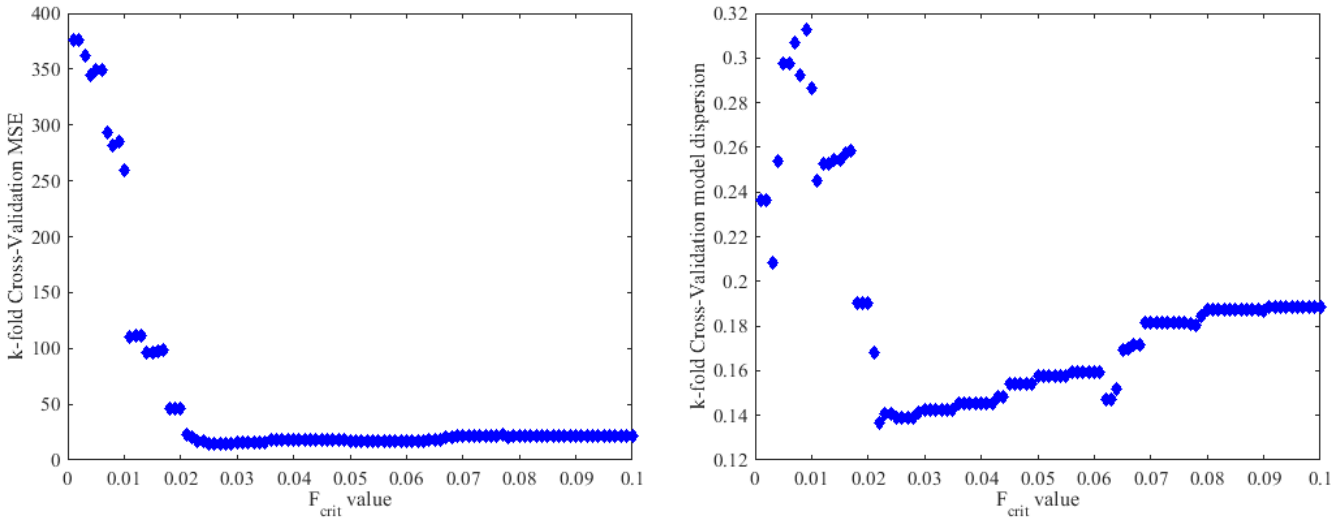


Figure 5-3: β values depending on chosen F_{crit} value for stepwise regression algorithm

In this picture, lines with marker represent β coefficients of predictor that have been used to generate synthetic data. Other β coefficients are represented with black line. Dotted and solid lines with markers represent the true and extracted β values respectively. We see that for small values of F_{crit} (below 0.07) only relevant predictors are selected. Thus F_{crit} should lies in the interval]0;0.07]. If F_{crit} is above 0.07 the method starts to extract non zero coefficient for predictor that are not in the model. Using a suitable value for F_{crit} doesn't lead to extract the correct β coefficients since we see discrepancies between doted and solid lines. This is due the artificial noise added to the response before performing stepwise regression. Bias in β values due to noise depends on the noise level and the model itself. Figure 5-4 shows the k-fold CV MSE and model variance depending on parameter F_{crit}.

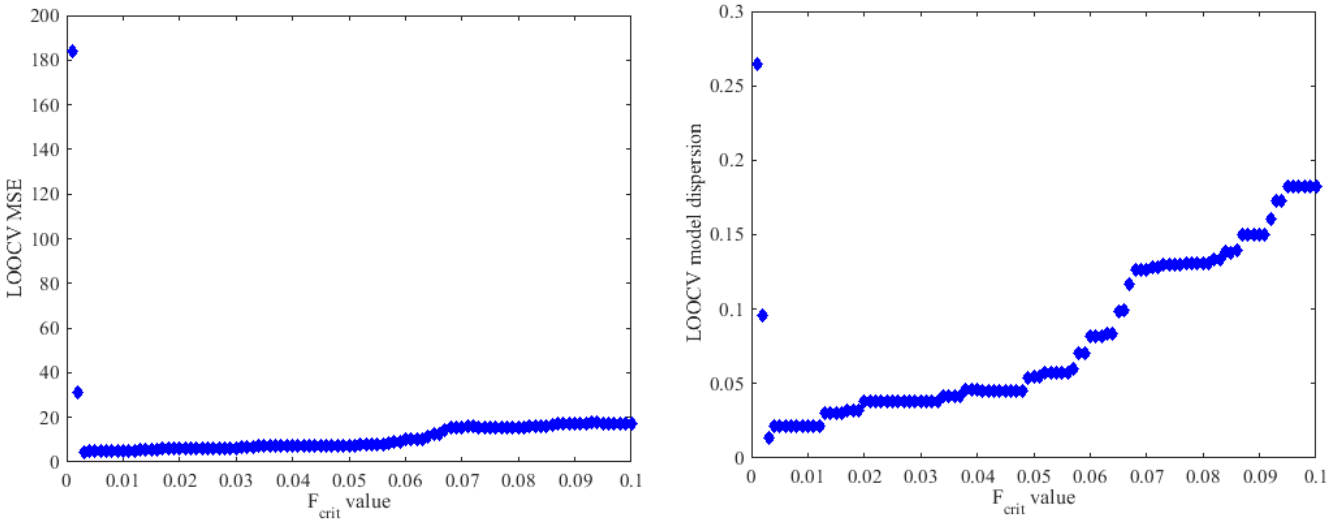


(a) (b)

Figure 5-4: MSE (a) and model variance (b) extracted using k-fold CV test

K-fold CV test shows an optimal point for $F_{crit} \approx 0.023$. This optimal point minimizes both model variance and MSE. This is a good result in view of Figure 5-3. K-fold CV test shows also that using $F_{crit} \ll 0.023$ increases the average MSE and model variance. This implies that the method fails more often in distinguishing relevant from irrelevant predictors. The result is more dependent on the training test used. If $F_{crit} \gg 0.023$ (especially if $F_{crit} > 0.07$), MSE does not rise significantly but model error does. It suggests that relevant predictors are included in the model but as F_{crit} rises, the probability to include an irrelevant predictor in the model rises as well, leading to overfit.

Figure 5-5 shows the LOOCV MSE and model variance depending on parameter F_{crit}.

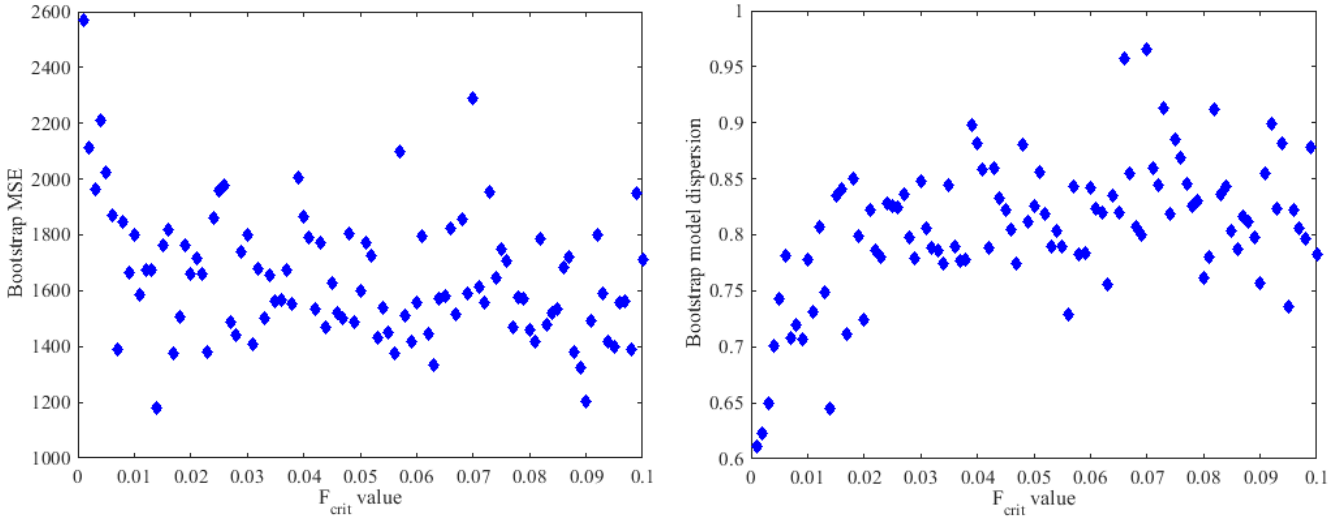


(a) (b)

Figure 5-5: MSE (a) and model variance (b) extracted using LOOCV test

Optimal value for F_{crit} is found to be 0.003, minimizing both MSE and model variance. This value lies within the correct range of F_{crit}. The same interpretations as for k-fold CV holds for LOOCV. The shift in the minimum of F_{crit} between these two methods is only due to the respective size of training and test dataset used. Both of these results are relevant.

Figure 5-6 shows the bootstrap MSE and model variance depending on parameter F_{crit}.



(a) (b)

Figure 5-6: Model variance against F_{crit} using bootstrap method

It can be noticed that bootstrap yields noisier results that LOOCV and k-fold CV. This can be explained by the fact that bootstrap method randomly draws the bootstrap samples. Indeed, there are too much possible bootstrap samples to compute all of them within a reasonable amount of time. With 25 observations, the number of different bootstrap combinations is $2.48 \cdot 10^{14}$. We have limited the computation to 150 bootstrap samples. However this noise also comes from the fact that stepwise regression is a bit cumbersome as a method, making it strongly dependent on the training set used to build the model. Thus for this method bootstrap is not suited to determine F_{crit} value.

To conclude, stepwise regression has been successfully calibrated using LOOCV, k-fold CV and bootstrap. Using the calibration, stepwise regression found the five relevant predictors among the 50 ones with only 25 observations impacted by 10% of noise.

5.3.2.5 Conclusion about stepwise regression

Compared to best subset selection, stepwise regression offers an easy way to find a good model within a reasonable amount of time. Indeed if the final model totalizes p-1 predictors among the p available, then the number of model to be calculated is:

$$N_{model} = 1 + \frac{p(p+1)}{2} \tag{151}$$

Thus for p=50, the algorithm will test 1327 models instead of $2.252 \cdot 10^{15}$ if we used best subset selection.

The main drawback of this method is that the final model is not guaranteed to be optimal in any specified sense. Moreover the procedure yields a single final model, although in practice there are often several equally good models. Many alternatives have been proposed involving a mixture of forward and backward stepwise regression but there are no convincing solutions since they yield different results without bringing more confidence on the result accuracy. Moreover this technique has been highly criticized in literature [166][167]. Thus this method should be used with care.

5.3.3 Shrinkage method [162]

As explained in previous paragraph, subset selection has drawbacks. In order to provide more robust methods, another approach called shrinkage method has been developed. Instead of adding or withdrawing successively variables from the model, this approach gradually reduces model coefficients value of least significant predictors. The approach is more robust because the results does not depends on any strategy chosen for the method (unlike stepwise regression) and is fast computed since only a limited amount of model construction is needed (unlike best subset selection).

5.3.3.1 Ridge regression [162]

Ridge regression has been introduced by A. N. Tikhonov **[168]**. This method is directly derived from ordinary least square regression. In least square regression, the solution minimizes the sum of square error between model and observations, that is, it finds the set of predictor coefficients that minimizes SSE, recalled here for convenience:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \tag{152}$$

In above equation, x are the predictors. β_0 is the intercept.

To that equation, ridge regression adds a second constrain that forces the β coefficients to be as small as possible. This constraint is introduced as a penalty called shrinkage penalty, calculated as the sum of square β times a constant λ that is set by the user. The ridge regression coefficient estimates β^R are the values that minimize

$$SSE + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{153}$$

Thus, if λ is set to 0, ridge regression is perfectly equivalent to ordinary least square. As λ increases, coefficients reduce and ultimately, if $\lambda = \infty$, then $\beta^R = 0$. Since equation (153) is no more a linear problem, it is usually minimized using Levenberg-Markard algorithm **[169][170]**.

This method is probably one of the most widely used for ill-posed problems. However, even if irrelevant predictor coefficients are quickly shrunk toward zero, these values cannot reach exactly zero. Thus it does not perform a practical variable selection, thus its investigation is out of the scope of this work.

5.3.3.2 Least Absolute Shrinkage and Selection method (LASSO)

This method has been introduced by R. Tibshirani **[171]**. It proposes an alternative to ridge regression that enables setting irrelevant predictor coefficient to exactly zero. The principle of LASSO is the same as ridge. The difference lies in the equation to be minimized. In this method the penalty is calculated using L^1 norm instead of L^2 norm of predictor coefficient. The function to be minimized becomes then:

$$SSE + \lambda \sum_{j=1}^{p} |\beta_j| \tag{154}$$

Again, in this method, λ coefficient has to be chosen by the user. Typically this parameter is determined using cross-validation or bootstrap methods.

In this work we used the Matlab implemented LASSO method **[172][173]**. In order to illustrate the behavior of this method, Figure 5-7 shows the estimated predictor coefficient β depending on parameter λ.

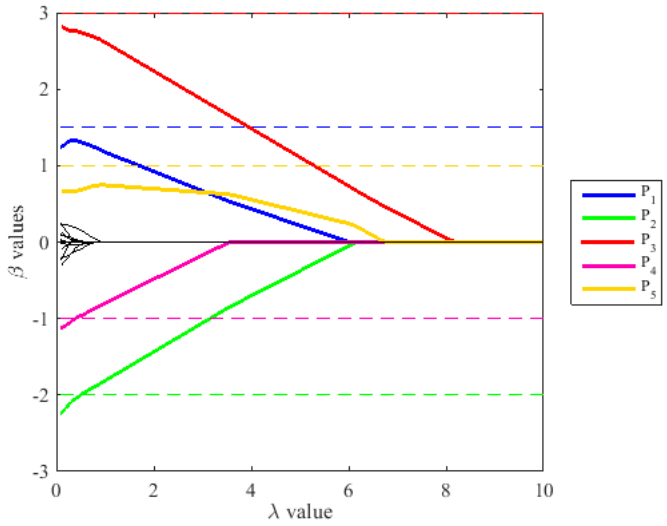
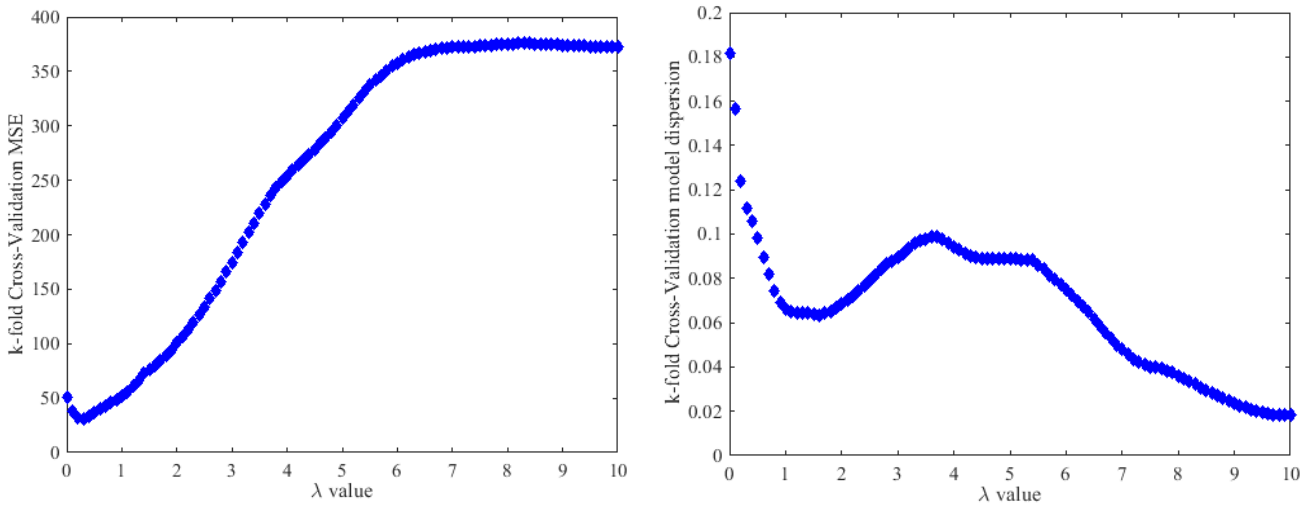


Figure 5-7: β values depending on chosen λ value for LASSO algorithm.

Using LASSO, the number of eliminated predictor is proportional to λ. We see that for $1 < \lambda < 3.5$, the model succeed to only select the relevant predictors. However including λ also biases the extraction of β coefficient, reducing their values. Thus, the method provides a powerful way to select

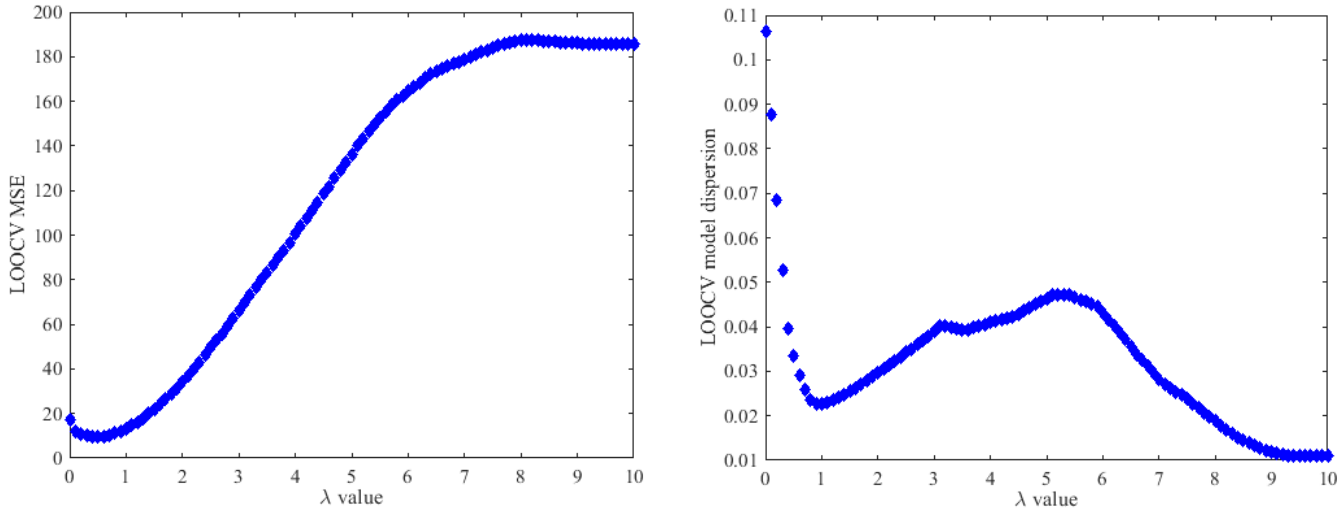
predictor but λ coefficient should be carefully chosen in order to optimize variable selection and minimize coefficient bias.



(a) (b)

Figure 5-8: MSE (a) and model variance (b) extracted using k-fold CV test

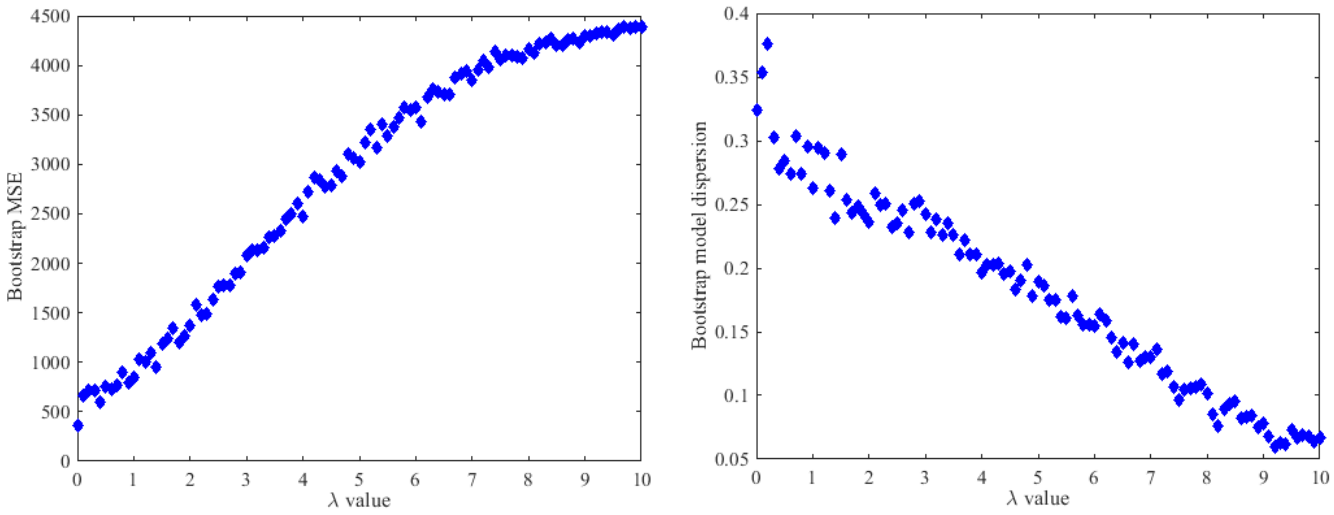
Figure 5-8 shows the k-fold CV MSE and model variance depending on parameter λ. This figure shows that a good tradeoff between MSE and model variance leads to an optimal λ value of 1.5 that a correct value since it would lead to select only relevant parameters.



(a) (b)

Figure 5-9: MSE (a) and model variance (b) extracted using LOOCV test

Figure 5-9 shows the LOOCV MSE and model variance depending on parameter λ. In the same trend that k-fold CV, LOOCV shows that a good tradeoff between model variance and MSE would lead to an optimal λ value of 1. It is slightly low and choosing this value might lead to a model that includes irrelevant predictor (depending on the training set) but anyway, there will be only few of them and their corresponding β coefficient would be very low compared to the others.



(a) (b)

Figure 5-10: Model variance against F_{crit} using bootstrap method.

Figure 5-10 shows the MSE and model variance calculated bootstrap depending on parameter λ. Again, the optimal λ can be deduced from a tradeoff between MSE and model variance. However this time, the optimum is not obvious and strongly depends on the chosen tradeoff. Thus in this case k-fold CV and LOOCV would be preferred over bootstrap.

To conclude, it has been possible, using LASSO, k-fold CV and LOOCV, to extract only the 5 relevant predictors among the 50 ones. Using the average of the optimal λ value found using k-fold CV and LOOCV leads to a reasonable choice.

5.3.4 Hybrid approaches

The two previous approaches that are subset selection and shrinkage method are very different at first glance but there is actually a bond between them. Some hybrid algorithm have been developed that are able (with a slight modification) to perform both these approaches. In this paragraph we introduce two of them called stagewise and least angle regression.

5.3.4.1 Forward stagewise regression **[174]**

Forward stagewise regression is based on the same principle as forward stepwise regression. It starts with all coefficients equal to zero, and iteratively updates the coefficient of the variable that achieves the maximal absolute inner product with the current residual by a small amount ϵ. This variable is called "best candidate". Thus the main difference with stepwise regression is that, at each step, variable coefficients that are in the model are not calculated by OLS but instead the method only increases one coefficient (making the approach continuous). It makes the approach less cumbersome and avoids biased decision about variable inclusion and deletion. This procedure has an interesting connection to the LASSO: under some conditions, it is known that the sequence of forward stagewise estimates exactly coincide with the lasso path, as the step size ϵ goes to zero. This method, also being more robust than stepwise regression, needs much more step to yield the final results. Moreover since it is identical to LASSO method if $\epsilon \approx 0$, advantages of this approach are limited, thus its investigation is out of the scope of this thesis.

5.3.4.2 Least angle regression **[175]**

Least Angle Regression (LARS, S suggesting LASSO) is based on forward stagewise regression. In stagewise regression the ϵ amount to be added to the best candidate is fixed. It is necessary to fix ϵ small in order to detect precisely when the best candidate changes. In LARS algorithm, the sum of increment that should be added to the best candidate, before the best candidate shifts, is calculated at once. This strongly reduces the computation time and makes it as efficient as stepwise regression. Applying a simple modification of the algorithm enables LARS to perform either stagewise regression or LASSO. LARS is thus an intermediate approach between LASSO and stagewise regression.

In this work we used the Matlab implemented LARS method **[175][176]** by Xiaohui Chen. In order to illustrate the behavior of this method, Figure 5-11 shows the estimated predictor coefficient β depending on the number of step λ.

Using LARS, the more steps the algorithm makes, the more predictors enter the model. If not stopping criterion is set, then the algorithm yield a model that comprises every predictors. Hopefully it is possible to set a maximum limit for the L_1 norm of predictor coefficients. This criterion is the calibration parameter and it is quite similar to the inverse of LARS λ parameter. Figure 5-11 shows that the L_1 norm of predictor coefficients is a critical parameter and should be calibrated in order to determine which model comprises only the relevant predictors. In this particular case, a correct number of steps (that leads to only select relevant predictors) should be comprised between 13 and 22. In this case even if the correct predictors are selected to enter the model, we see that their coefficients are far from the exact value. This is due to the "least angle approach" that does not calculate β using least square fit but instead it increments them gradually. However it is suited to perform variable selection. β can then be calculated using least square fit after variable selection.

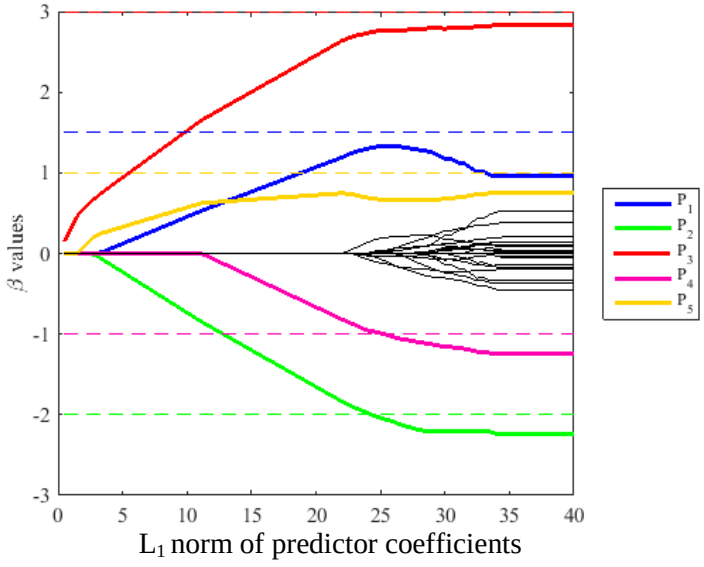


Figure 5-11: β values depending on chosen L_1 norm value for LARS algorithm.

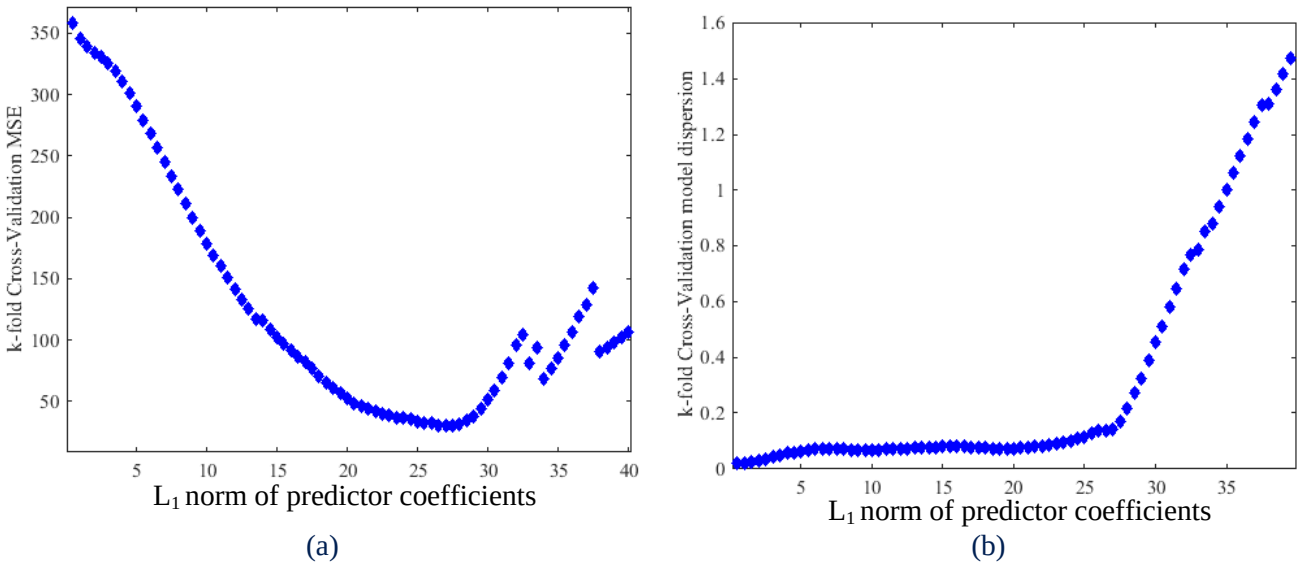


Figure 5-12: MSE (a) and model variance (b) extracted using k-fold CV test

Figure 5-12 shows the k-fold CV MSE and model variance depending on number of step λ. From this figure, we see that enabling more or less 25 steps leads to a good trade of between MSE and model variability. It is a bit too much in order to select only relevant predictors. But at this step, the selected irrelevant predictors have very small coefficient. Thus their impact in the model is limited.

Figure 5-13 shows the LOOCV MSE and model variance depending on number the L_1 norm of predictor coefficients. Using LOOCV the same kind of trend is obtained compared to k-fold CV method. Here the optimal value is rather around 22, what is satisfying since it would lead to select only relevant predictors.

Figure 5-14 shows the bootstrap calculated model variance depending on the L_1 norm of predictor coefficients. This figure shows that bootstrap can be used to determine the optimal the L_1 norm value through a tradeoff between MSE and model dispersion. However, as in the case of LASSO, the results will strongly depend on the chosen tradeoff. Thus it is safer to rely on k-fold CV and LOOCV.

To conclude about LARS algorithm, it has been successfully applied to extract the relevant predictors as LASSO and stepwise regression.

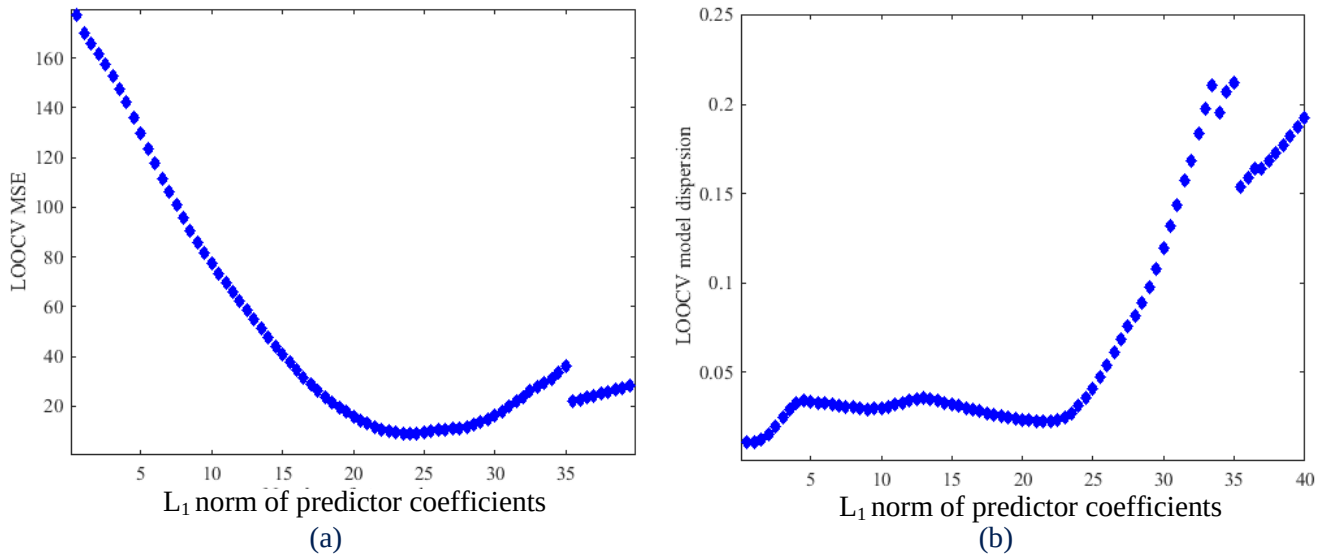


Figure 5-13: MSE (a) and model variance (b) extracted using LOOCV test.

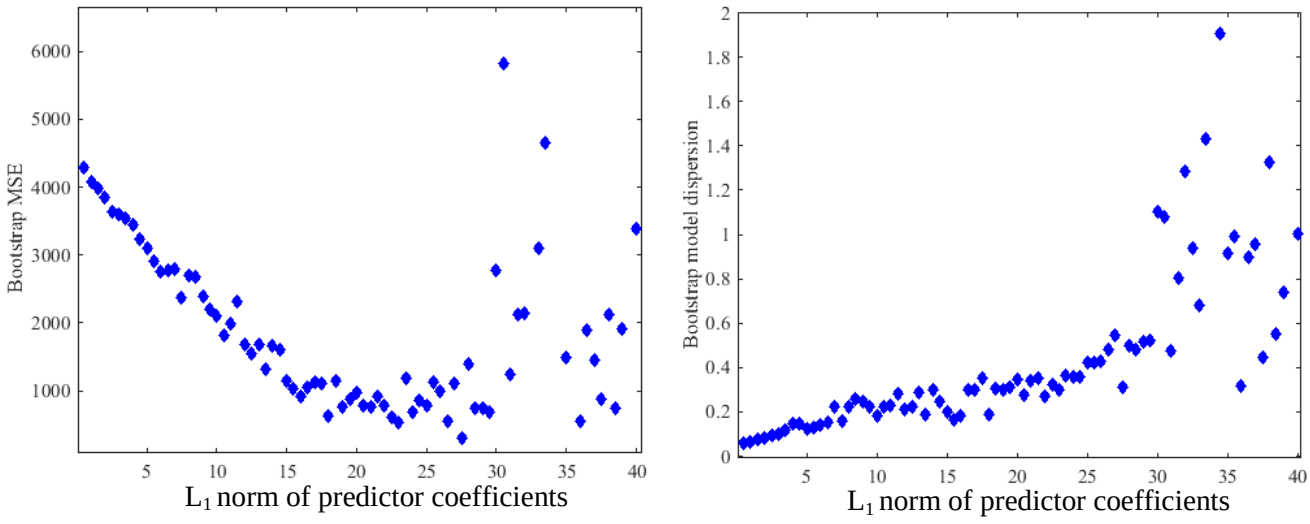


Figure 5-14: Model error (a) and variance (b) against L_1 norme using bootstrap method.

5.3.5 Conclusion about variable selection methods

In this paragraph we have introduced and tested three methods to select relevant variables among a large number of irrelevant one using a limited amount of noisy observations. Application showed that k-fold CV and LOOCV method have been able to calibrate parameters for each method. The 5 relevant predictors used in the model to compute synthetized data have been discriminated by the three methods among 50 variables using 25 observations impacted by 10% of artificial noise. This application addresses the 2 main issues that one can face when building a PCM: ill-posed problem and noise in the observable. It should be noted that even though predictor values have been randomly drawn, the limited number of observation leads to correlations between some predictors (up to 0.51 for the correlation factor between predictor P_4 and P_{49}). Thus the three models also deal with moderately correlated predictors.

Every PCM construction method works fine in this case but the cross-validation methods can lead to slightly different results. The best practice is to use the three cross-validation tests to calibrate the PCM construction method, comparing their results. For example if k-fold CV test suggests the same calibration than LOOCV but a different calibration compared to bootstrap, then results drawn from k-fold CV and LOOCV can be considered more reliable and should be used instead of bootstrap results. Considering PCM construction methods, there is no general rule in order to decide which one to use. However it can be noticed that stepwise is the fastest, followed by LASSO and then LARS. However, stepwise regression will fail more easily than LASSO and LARS if a large amount of noise is considered. Moreover LARS and LASSO give a continuous trend of β against the calibration parameter. Thus it is possible to rank predictors in term of relevance. This is not possible with stepwise regression since any predictor can enter or leave the model at any F_{crit} increment. In practice every PCM construction method should be tested in order to find the best model.

5.4 Application to TCAD simulations

In previous paragraphs we have introduced all the required tools on order to model build PCM according to Figure 5-1. This paragraph aims at demonstrating that these tools are required in order to build robust PCM. Indeed, the strategy is to use a two-stage PCM, as it has been shown in the introduction (PCM scheme is recall in Figure 5-15 for convenience).

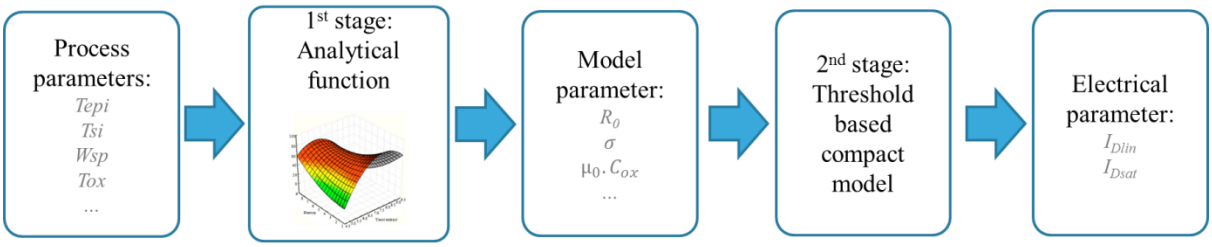


Figure 5-15: Scheme of the two-stage PCM

First, one could think of a more simple approach that directly relates process and electrical parameters without using a compact model (avoiding stage two). We will show that this simplification is less advantageous than the two-stage PCM by comparing them in §5.4.1 and §5.4.2. In particular we will show that building PCM that directly links electrical and process parameters is not handy and efficient. Using the two-stage PCM in §5.4.2, we will be able to model the linear and saturation drain current over the whole strong inversion range of V_G. In these two paragraph, every PCM will constructed using TCAD simulations DOE introduced in §3.4 and Ordinary Least Square (OLS). Moreover, within wafer variability, local variability and noise will not be considered.

Thereafter will demonstrate that, in practical case, using OLS is not efficient at all and variable selection methods (stepwise regression, LASSO and LARS) should be used instead with the PCM construction flow as depicted in Figure 5-1. This will be shown with a set of TCAD simulations that mimics silicon measurements at die level across a wafer, including within-wafer process variability. PCM construction using OLS with these data will not work at all since it will not be able to select variables. Variable selection methods will be applied afterward on the same dataset, building proper PCM successfully.

5.4.1 Building TCAD simulated I_{Dlin} and I_{Dsat} PCM using OLS

In this paragraph, we investigate the simple approach that directly relates process and electrical parameters without using a compact model. We build a PCM for I_{Dlin} and I_{Dsat} using OLS and the TCAD simulated DOE introduced in chapter 3. The DOE is a factorial design. Details about the DOE are recalled here for convenience, listing each process parameter included in the DOE and their related experimental values:

- Epitaxial height (Tepi) [12, 14, 16] nm
- Channel thickness (Tsi) [5, 6, 6.6, 8] nm
- Spacer width (Wsp) [8, 10.35, 12] nm
- Implanted dopant dose (f dose) [0.5, 0.7, 1, 1.2, 1.5] (All source-drain and LDD implant dose are multiplied by this factor)
- Insulating layer (IL) thickness (T_{il}) [0.8, 1.05, 1.2, 1.8, 2.5, 4] nm
- IL/High K interfacial charges (Qhk) [10^{10}, 10^{11}, 10^{12}, 3.10^{12}, 10^{13}] cm^{-2}
- Contact resistance (Rext) [20, reference, 200, 500] Ω (Reference values are 90 and 212 Ω for nMOS and pMOS respectively)
- Spike anneal (Tspike) [800, 1000, 1052, 1100]

To this factorial design we have added some cross terms. Corresponding experiments are detailed in Table 5-1. These experiences are referenced as “Mixed” in Figure 5-17. This simulation setup does neither account for local or within-wafer variability nor for measurement noise. In order to model I_{Dlin} and I_{Dsat} dependence on process parameters we built their PCM upon simulation results using OLS. OLS only accounts for process parameters included in the DOE. Since there is no variability, other

process parameters are fixed and thus cannot play any role in I_{Dlin} and I_{Dsat} variations. Of course, in practical cases, these assumptions do not hold. We shall see later the consequences.

	Process parameters	
Experience	Wsp (nm)	Tsi (nm)
1	8	5
2	8	8
3	12	5
4	12	8

Table 5-1: Cross terms experiences simulated by TCAD

Figure 5-16 shows the relations between process and electrical parameters depending if we are constructing the PCM or if we exploit it.

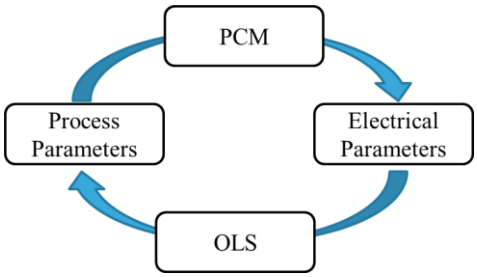
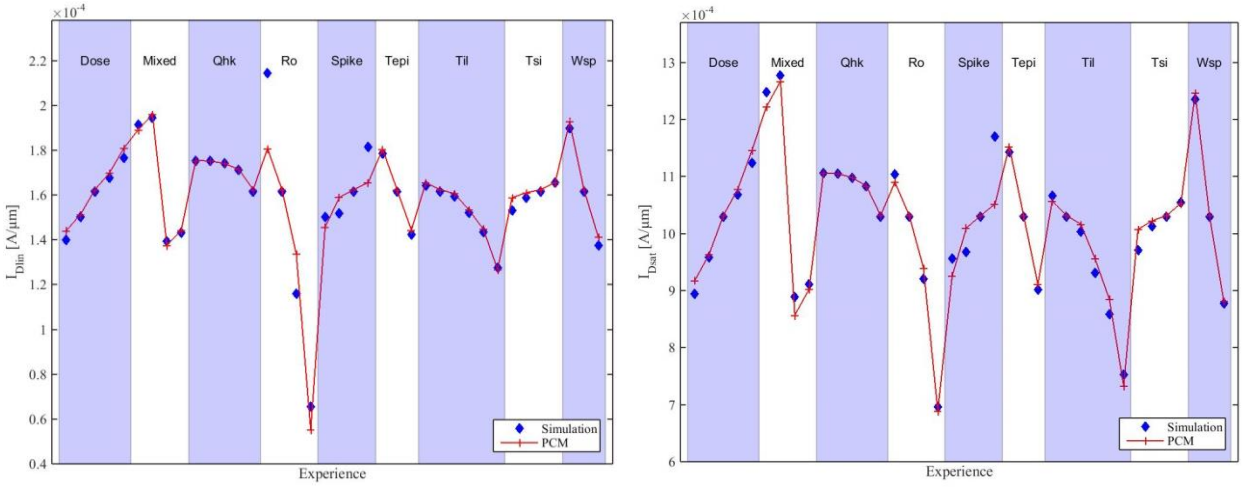


Figure 5-16: Flow chart of process and electrical relations

This figure shows that, in the current application, we directly relate process parameters thanks to polynomial formula. These polynomial formula are constructed using OLS.



(a) (b)

Figure 5-17: I_{Dlin} and I_{Dsat} simulated and model using OLS for transistor of nominal gate length with $V_G=V_{DD}$.

I_{Dlin} and I_{Dsat} are sensitive to every process parameter included in the DOE. Thus, their PCM account for all of them. In Figure 5-17, results shows that the PCM is quite accurate although some process parameter dependences are not linear in the considered range (especially for R_{ext} Tspike and Tsi). However this approach is not satisfying since it provides a PCM that is suited for only one gate and drain bias and one gate length. Thus, modeling the whole strong inversion regime in linear and saturation regime for every channel length would require a large number of PCM, making the global model discontinuous and not easy to handle. I_{Dlin} and I_{Dsat} PCM coefficients are gathered in Table 5-2.

Associated predictor with coefficient unit	I_{Dlin} Coefficients	I_{Dsat} Coefficients
Constant [$\mu A/\mu m$]	0.35	2.30
$Tepi$ [$\mu A/\mu m^2$]	-9.03	-60.4
Wsp [$\mu A/\mu m^2$]	-12.9	-91.2
Tsi [$\mu A/\mu m^2$]	2.27	15.2
Til [$\mu A/\mu m^2$]	-12.1	-101
f_dose [µA/µm]	$3.68\cdot 10^{-2}$	0.228
Tspike [µA/µm/°C]	$6.63\cdot 10^{-5}$	$4.22\cdot 10^{-4}$
Qhk [µA.µm]	$-1.29\cdot 10^{-7}$	$-7.41\cdot 10^{-7}$
$Rext$ [$\mu A/\mu m/\Omega$]	$-2.62\cdot 10^{-4}$	$-8.37\cdot 10^{-4}$

Table 5-2: I_{Dlin} and I_{Dsat} PCM coefficients.

5.4.2 Building PCM for TCAD simulated model parameters using OLS

A solution to the issue raised in previous paragraph is to build a two-stage PCM as introduced in Figure 5-1. This alternative is investigated here, where analytical model is created for the model parameters introduced in chapter 2 (such as R_0, σ, $\mu_0.C_{ox}$, θ_2, Vt_{lin}, Vt_{sat}, $v^*.C_{ox}$) instead of drain current. Modeling these parameters would enable modeling the whole strong inversion regime in saturation and linear regime using only 7 PCMs.

Figure 5-18 shows the relations between process and electrical parameters depending if we are constructing the PCM or if we exploit it.

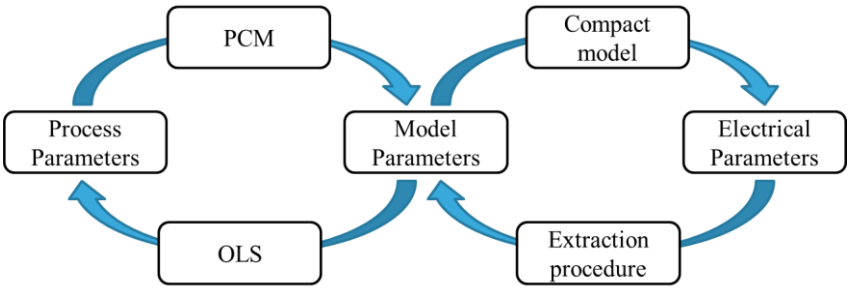
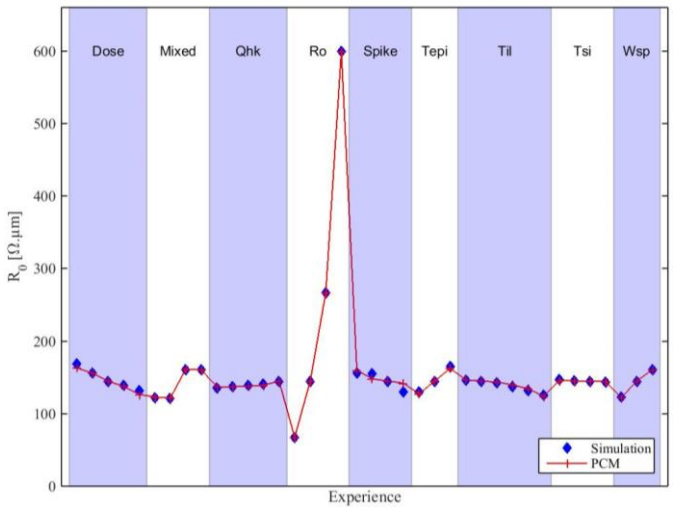


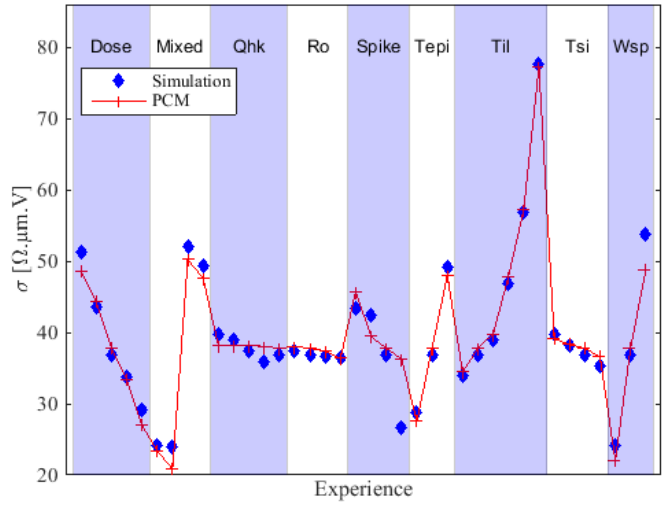
Figure 5-18: Flow chart of process and electrical relations

This figure is to be compared with Figure 5-16. It shows that, in this current application, we relate process parameters via the compact model instead of relating them directly as we did in previous paragraph. Figure 5-19 shows the PCM obtained using OLS for model parameters R_0, σ, $\mu_0.C_{ox}$, θ_2, Vt_{lin}, Vt_{sat} and $v^*.C_{ox}$.



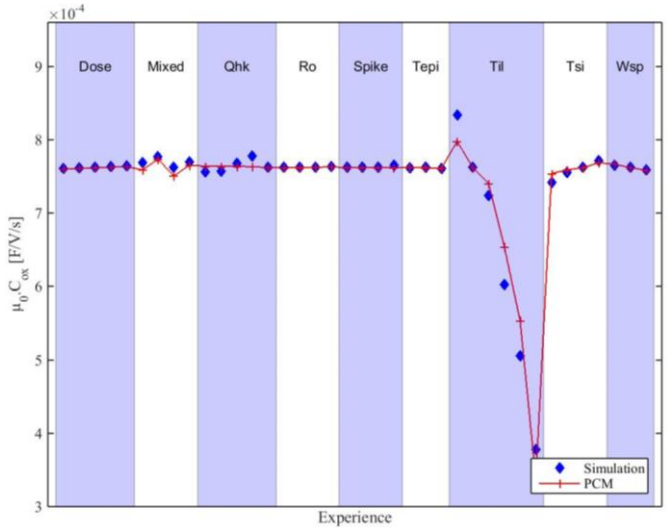
(a)

$R0\ [\Omega.\mu m]$	Coefficients
$Constant\ [\Omega.\mu m]$	$-7.65 \cdot 10^{1}$
$Tepi\ [\Omega]$	$8.79 \cdot 10^{3}$
$Wsp\ [\Omega]$	$9.70 \cdot 10^{3}$
$Tsi\ [\Omega]$	$-4.68 \cdot 10^{2}$
$Til\ [\Omega]$	$-7.13 \cdot 10^{3}$
$f\_dose\ [\Omega.\mu m]$	$-3.62 \cdot 10^{1}$
$Tspike\ [\Omega.\mu m/°C]$	$-5.99 \cdot 10^{-2}$
$Qhk\ [\Omega.\mu m\text{^}3]$	$7.84 \cdot 10^{-5}$
$Rext\ [\mu m]$	1.11



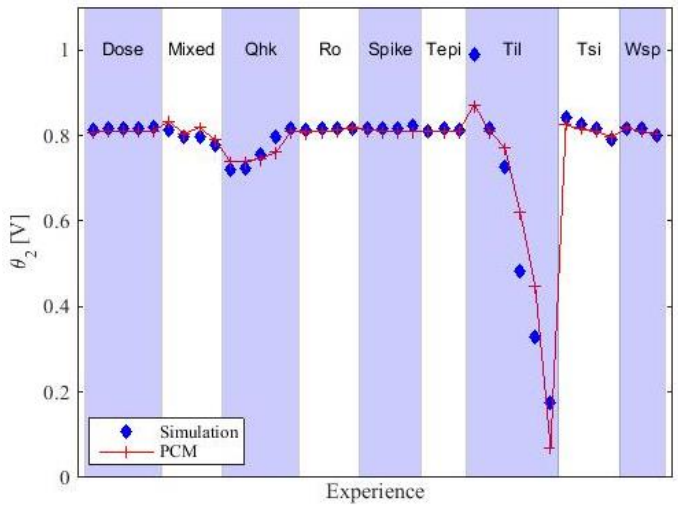
(b)

$\sigma\ [\Omega.\mu m.V]$	Coefficients
$Constant\ \ [\Omega.\mu m.V]$	$-5.58 \cdot 10^{1}$
$Tepi\ [\Omega.V]$	$5.11 \cdot$
$Wsp\ [\Omega.V]$	$6.70 \cdot 10^{3}$
$Tsi\ [\Omega.V]$	$-8.56 \cdot 10^{2}$
$Til\ [\Omega.V]$	$1.34 \cdot 10^{4}$
$f_{dose}\ \ [\Omega.\mu m.V]$	$-2.17 \cdot 10^{1}$
$Tspike\ [\Omega.\mu m.V/°C]$	$-3.17 \cdot 10^{-2}$
$Qhk\ [\Omega.\mu m^3.V]$	$-4.01 \cdot 10^{-6}$
$Rext\ [\mu m.V]$	$-3.49 \cdot 10^{-3}$



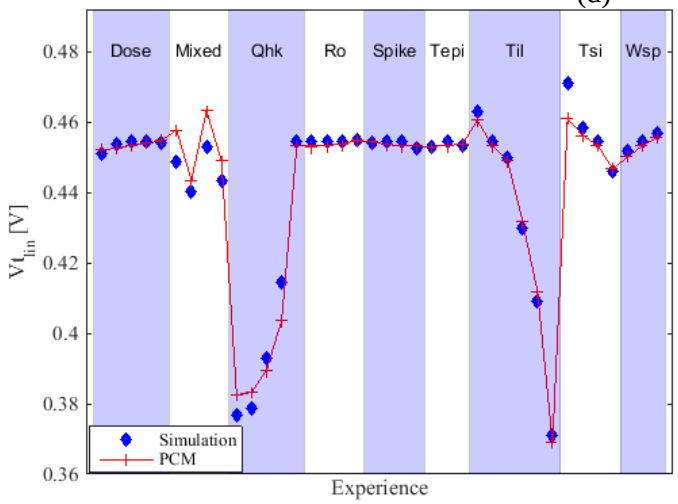
(c)

$\mu_0.C_{ox}\ [F/V/s]$	Coefficients
$Constant\ \ [F/V/s]$	$9.06 \cdot 10^{-4}$
$Tepi\ [F/V/s/\mu m]$	$-3.97 \cdot 10^{-4}$
$Wsp\ [F/V/s/\mu m]$	$-2.04 \cdot 10^{-3}$
$Tsi\ [F/V/s/\mu m]$	$4.94 \cdot 10^{-3}$
$Til\ [F/V/s/\mu m]$	$-1.44 \cdot 10^{-1}$
$f_{dose}\ \ [F/V/s]$	$3.23 \cdot 10^{-6}$
$Tspike\ [F/V/s/°C]$	$-1.75 \cdot 10^{-11}$
$Qhk\ [F.\mu m^2/V/s]$	$-2.09 \cdot 10^{-11}$
$Rext\ [F/V/s/\Omega]$	$3.13 \cdot 10^{-9}$



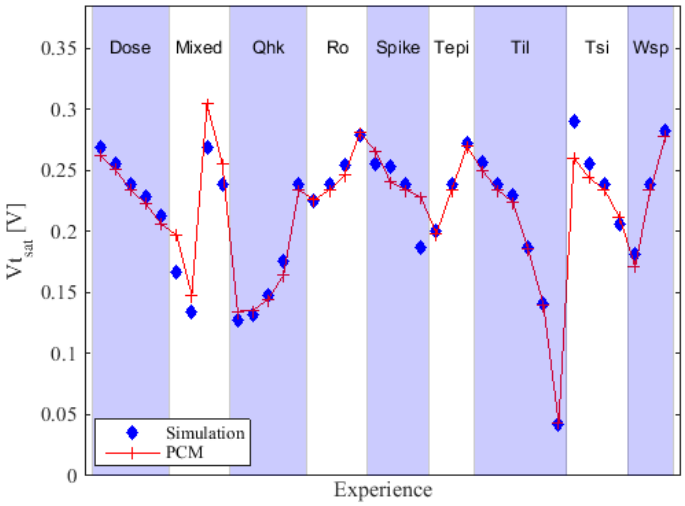
(d)

$\theta_2\ [V^{-2}]$	Coefficients
$Constant\ [V^{-2}]$	1.10
$Tepi\ [V^{-2}/\mu m]$	$6.23 \cdot 10^{-1}$
$Wsp\ [V^{-2}/\mu m]$	-3.59
$Tsi\ [V^{-2}/\mu m]$	-9.31
$Til\ [V^{-2}/\mu m]$	$-2.51 \cdot 10^{2}$
$f_{dose}\ [V^{-2}]$	$2.99 \cdot 10^{-3}$
$Tspike\ [V^{-2}/°C]$	$-1.52 \cdot 10^{-5}$
$Qhk\ [V^{-2}.\mu m^2]$	$7.03 \cdot 10^{-7}$
$Rext\ [V^{-2}/\Omega]$	$2.15 \cdot 10^{-5}$



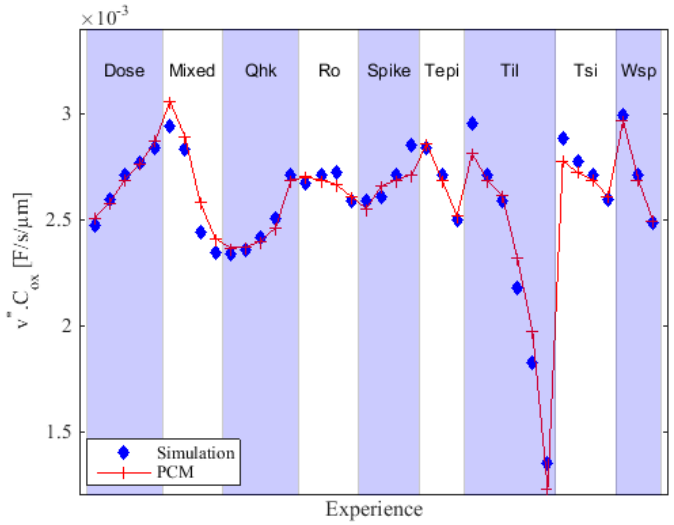
(e)

$Vt_{lin}\ [V]$	Coefficients
$Constant\ [V]$	$4.29 \cdot 10^{-1}$
$Tepi\ [V/\mu m]$	$1.58 \cdot 10^{-1}$
$Wsp\ [V/\mu m]$	1.37
$Tsi\ [V/\mu m]$	-4.74
$Til\ [V/\mu m]$	$-2.86 \cdot 10^{1}$
$f_{dose}\ [V]$	$2.69 \cdot 10^{-3}$
$Tspike\ [V/°C]$	$-3.99 \cdot 10^{-6}$
$Qhk\ [V.\mu m^2]$	$7.09 \cdot 10^{-7}$
$Rext\ [V/\Omega]$	$3.48 \cdot 10^{-6}$



(f)

$Vt_{sat}\ [V]$	Coefficients
$Constant\ [V]$	$-3.63 \cdot 10^{-2}$
$Tepi\ [V/\mu m]$	$1.78 \cdot 10^{1}$
$Wsp\ [V/\mu m]$	$2.68 \cdot 10^{1}$
$Tsi\ [V/\mu m]$	$-1.64 \cdot 10^{1}$
$Til\ [V/\mu m]$	$-6.47 \cdot 10^{1}$
$f_{dose}\ [V]$	$-5.62 \cdot 10^{-2}$
$Tspike\ [V/°C]$	$-1.27 \cdot 10^{-4}$
$Qhk\ [V.\mu m^2]$	$9.95 \cdot 10^{-7}$
$Rext\ [V/\Omega]$	$1.14 \cdot 10^{-4}$



(g)

$v^*.C_{ox}\ [F/\mu m/s]$	Coefficients
$Constant\ [F/\mu m/s]$	$4.78 \cdot 10^{-3}$
$Tepi\ [F/\mu m^2/s]$	$-8.54 \cdot 10^{-2}$
$Wsp\ [F/\mu m^2/s]$	$-1.19 \cdot 10^{-1}$
$Tsi\ [F/\mu m^2/s]$	$-5.61 \cdot 10^{-2}$
$Til\ [F/\mu m^2/s]$	$-4.93 \cdot 10^{-1}$
$f_{dose}\ [F/\mu m/s]$	$3.63 \cdot 10^{-4}$
$Tspike\ [F/\mu m/s/°C]$	$5.36 \cdot 10^{-7}$
$Qhk\ [F.\mu m/s]$	$3.20 \cdot 10^{-9}$
$Rext\ [F/\mu m/s/\Omega]$	$-1.92 \cdot 10^{-7}$

Figure 5-19: PCM modeled and TCAD simulated R_0 (a), σ (b), $\mu_0.C_{ox}$ (c), θ_2 (d), Vt_{lin} (e), Vt_{sat} (f), $v^*.C_{ox}$ (g).

An overview of Figure 5-19 plots shows that PCMs are accurate. It can be noted that T_{il} and Tspike effects are not very well approximated with linear dependence. For example $v^*.C_{ox}$, θ_2, and $\mu_0.C_{ox}$ dependence on T_{il} is not very accurate. As well $v^*.C_{ox}$, Vt_{sat}, σ, and R_0 dependence on Tspike is not very accurate.

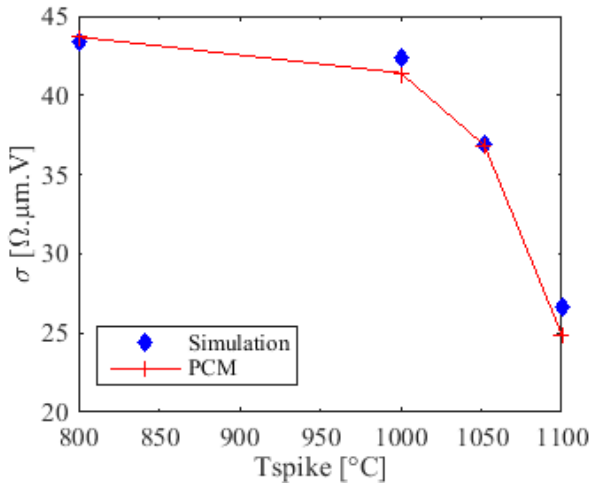
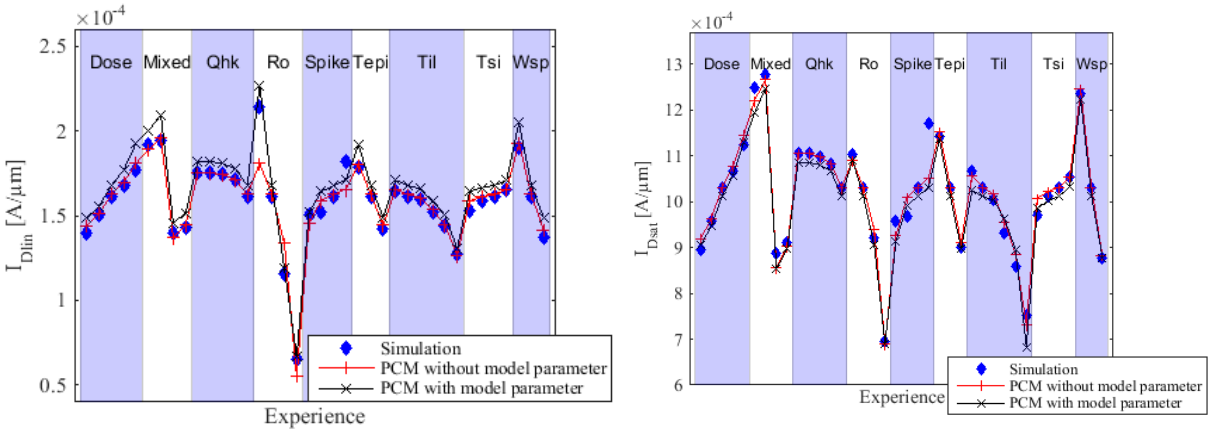


Figure 5-20: σ extracted from simulations and modeled depending on anneal temperature.

A closer look at the last example shows that model parameters have an exponential dependence on Tspike rather than a linear one. A empirical fit of σ against Tspike with an exponential formulation is shown in Figure 5-20.

$$\sigma = 43.7 - 1.75 \cdot 10^{-9} \cdot \exp(0.021 \cdot T_{spike}) \quad (155)$$

Thus PCM built using OLS are accurate except for modeling the impact of anneal temperature and insulator layer thickness. However, impact of these two process parameters on model parameters can be corrected using nonlinear empirical formula and variable shift (i.e. using $\exp(0.021.T_{spike})$ instead of T_{spike} as predictor for the PCM).



(a) (b)

Figure 5-21: Comparison between I_{Dlin} and I_{Dsat} model with PCM straightforwardly build upon I_{Dlin} and I_{Dsat} or build upon model parameters.

Figure 5-21 shows I_{Dlin} and I_{Dsat} modeled using OLS (same results have been shown in Figure 5-17) and modeled using the compact model and PCMs for model parameters. We see that both models are very close. I_{Dlin} PCM using the compact model is more accurate in the range of R_0. Thus using a two-stage PCM has two advantages: i) it is slightly more accurate, ii) it enable a continuous modeling of the full strong inversion regime in linear and strong inversion regime with only 7 PCM whereas a one stage PCM would require as much PCM as gate biases making the model cumbersome and discontinuous.

Even though, this simple approach works fine using TCAD simulations, it is not able to select variables. And the constructed PCMs suggest some unphysical relationships. In order to get a more reliable model, variable selection should be applied.

5.4.3 Building PCM in a silicon-like case, based on within wafer variability

In this section we demonstrate that OLS used in previous approach are not suited to build PCM. Indeed, it has already shown limitations, as mentioned in previously. In addition, we will show that this technique is even more limited if we want to apply it on silicon measurements. Indeed, went processing an experiment on silicon, we have to face within wafer variability. This means that every process parameters fluctuates more or less depending on the position on the wafer. This variability strongly impact the drain current as it has been shown in chapter 4, where wafer level drain current box plot displayed large dispersion. Hence, using OLS to build PCM with process and electrical parameters averaged over each wafer would lead to a large uncertainty in the resulting PCM. Another solution consists in monitoring parameters (electrical and process ones) at die level. The uncertainty about monitored parameters will thus be limited to local variability. This leads to consider every process parameter for PCM construction. Since the number of process parameters is large and only a limited part of them are actually relevant, variable selection should be made. This is what we demonstrate here, using TCAD simulations that mimics a wafer measured a die level. Simulations include within-wafer variability for process parameters. Process parameters statistics are shown in Table 5-3.

Process Parameters	Tepi [nm]	Wsp [nm]	Tsi [nm]	T_{il} [nm]	f_dose	Tspike [°C]	R_{ext} [Ω. µm]	Qhk
Average	14	10.35	7	1.2	1	1050	22.5	10^{12}
Standard deviation	1.3	1.3	0.5	0.25	0.24	2	2.25	$2 \cdot 10^{11}$

Table 5-3: Process parameters average and dispersion over the set of simulations

The simulation setup consists in 100 sites simulated with random variations of Tepi, Wsp, Tsi, T_{il}, fdose, Tspike, Rext and Qhk. Each site contains 5 channel lengths ranging from 30 nm up to 1 µm. Average values and standard deviations of process parameters are regrouped in Table 5-3.

Before building PCM, "fake" process parameters have been added to the predictor matrix in order to simulate a more realistic situation where a large number of process parameters are accounted for. Fake parameters represent process parameters that have no influence on drain current. Since they are as much process parameters (predictors) than observations, the problem becomes ill-posed. "Fake" process parameters are simply randomly generated predictors. They are in no way related with the observation, but chance correlations can occur between them.

We will first attempt to build PCM using OLS and show that the approach completely fails because of the issue mentioned before. Then we will build PCM using variable selection methods and show that the approach is more efficient.

5.4.3.1 Using OLS

PCM construction is first done using OLS, as we did in §5.4.2. The flow chart of process and electrical parameters relationship is shown in Figure 5-22.

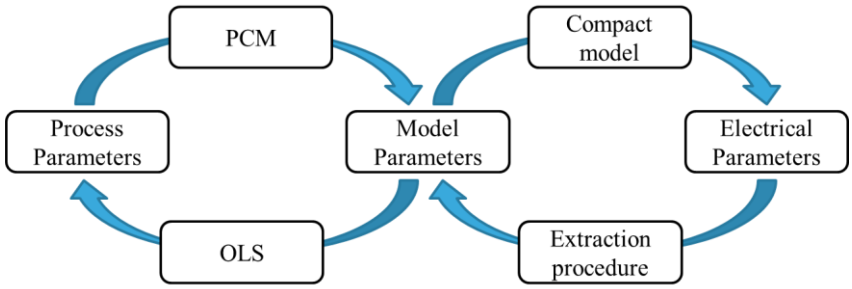
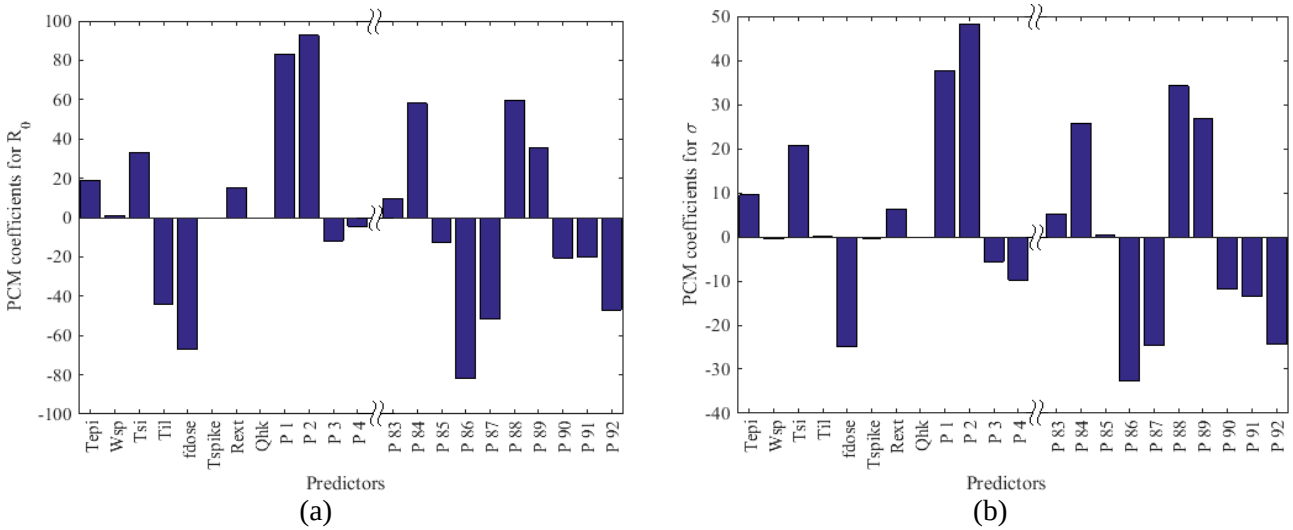


Figure 5-22: Flow chart of process and electrical relations

One PCM is built for each model parameter. Figure 5-23 shows the value of the different PCM coefficients for each PCM.



(a) (b)

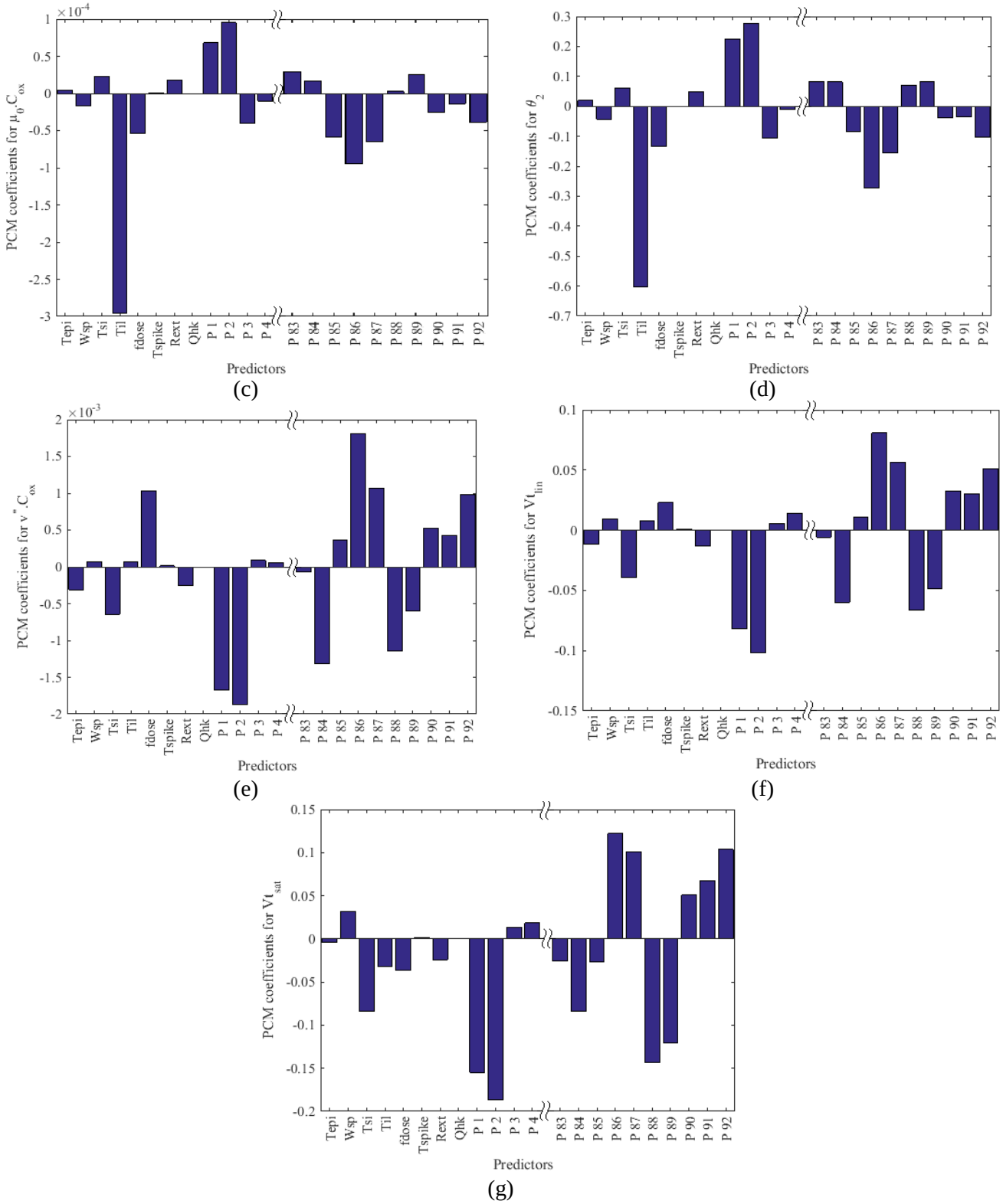


Figure 5-23: PCM coefficient of model parameters R_0 (a), σ (b), $\mu_0.C_{ox}$ (c), θ_2 (d), $v^*.C_{ox}$ (e), V_{tlin} (f) and V_{tsat} (g).

Figure 5-23 shows that using OLS, every predictor are included in the model. Moreover, fake predictors are accounted for with non-negligible coefficients. It means that in this situation OLS solution leads to overfit. The predictability and interpretability of such a model is very poor. Thus we cannot use OLS to exploit measurements at die scale, accounting for a large number of process parameters.

5.4.3.2 Using stepwise, LASSO and LARS

In response to OLS flaws, stepwise, LASSO and LARS methods are applied here to build model parameters PCM. Corresponding flow chart of process and electrical parameters relationships is shown in Figure 5-24.

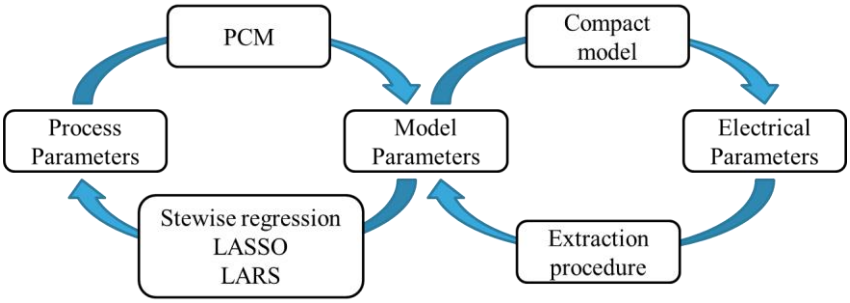


Figure 5-24: Flow chart of process and electrical relations

Compared to previous approach, here we use variable selection methods instead of OLS. Since these methods can handle ill-posed problem and perform variable selection, it will be more suited to build PCM upon measurements at die level including a large amount of process parameters.

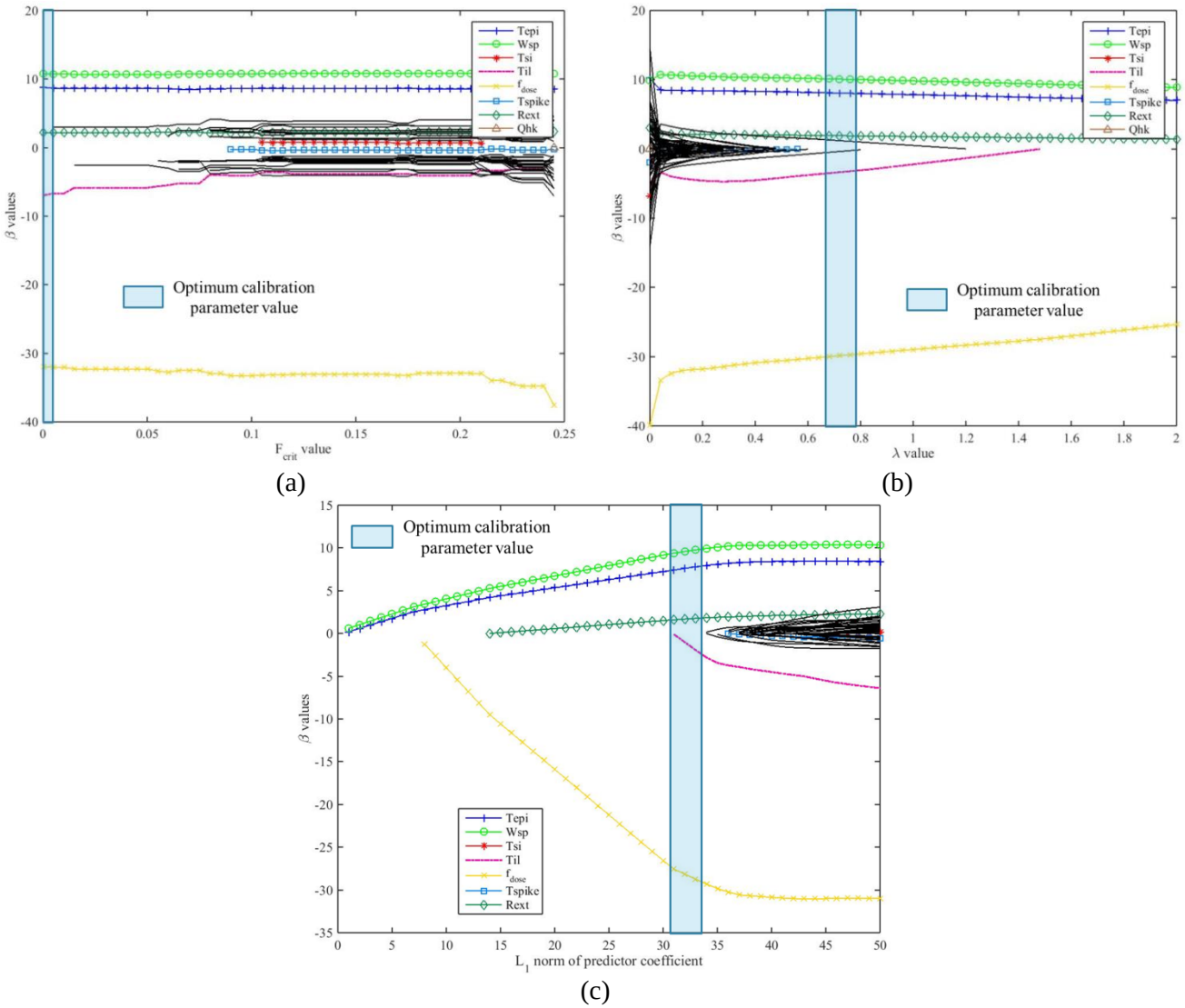


Figure 5-25: R_0 PCM constructed using stepwise regression (a), LASSO (b) and LARS (c) depending on their respective calibration parameter.

Figure 5-25 shows the result of R_0 PCM construction. Each subplot represents the results of one method against its calibration parameter. The optimum value of calibration parameter is indicated by the blue shaded area. This area has been determined using K-fold CV and LOOCV. Related plots are available in Appendix B. We see that the three methods have selected Wsp, Tepi, Rext, T_{il} and f_{dose} as predictors for R_0. LASSO has included also 2 non relevant predictors. So comparing the results of the three methods enable a clear distinction between the relevant and irrelevant predictors. It should be noted that Tspike does not appear in the model but T_{il} does. T_{il} is not physically related to R_0 and we expect it to not enter the model. However we have seen in Figure 5-19 (a) that T_{il} slightly impacts R_0 because of the flaws in the model related to its simplification (see chapter 3). The reason why T_{il} is accounted in the model and no Tspike is because of the relative dispersion of these parameters. Indeed T_{il} dispersion in these simulations represents 15.6% of the total dispersion simulated in TCAD DOE. This is large in comparison with Tspike dispersion that only represents 0.67% of Tspike total variation simulated in TCAD DOE.

Figure 5-26 shows the results of σ PCM construction. All 3 model includes T_{il}, Wsp, Tepi and fdose. This is in agreement with the PCM built using the TCAD simulated DOE. Again Tspike is missing in the model because of its low dispersion. LASSO method includes 3 extra 'fake' predictors to the model but their coefficients are very low compared to the others and these errors can be spotted by comparing this result with the other methods.

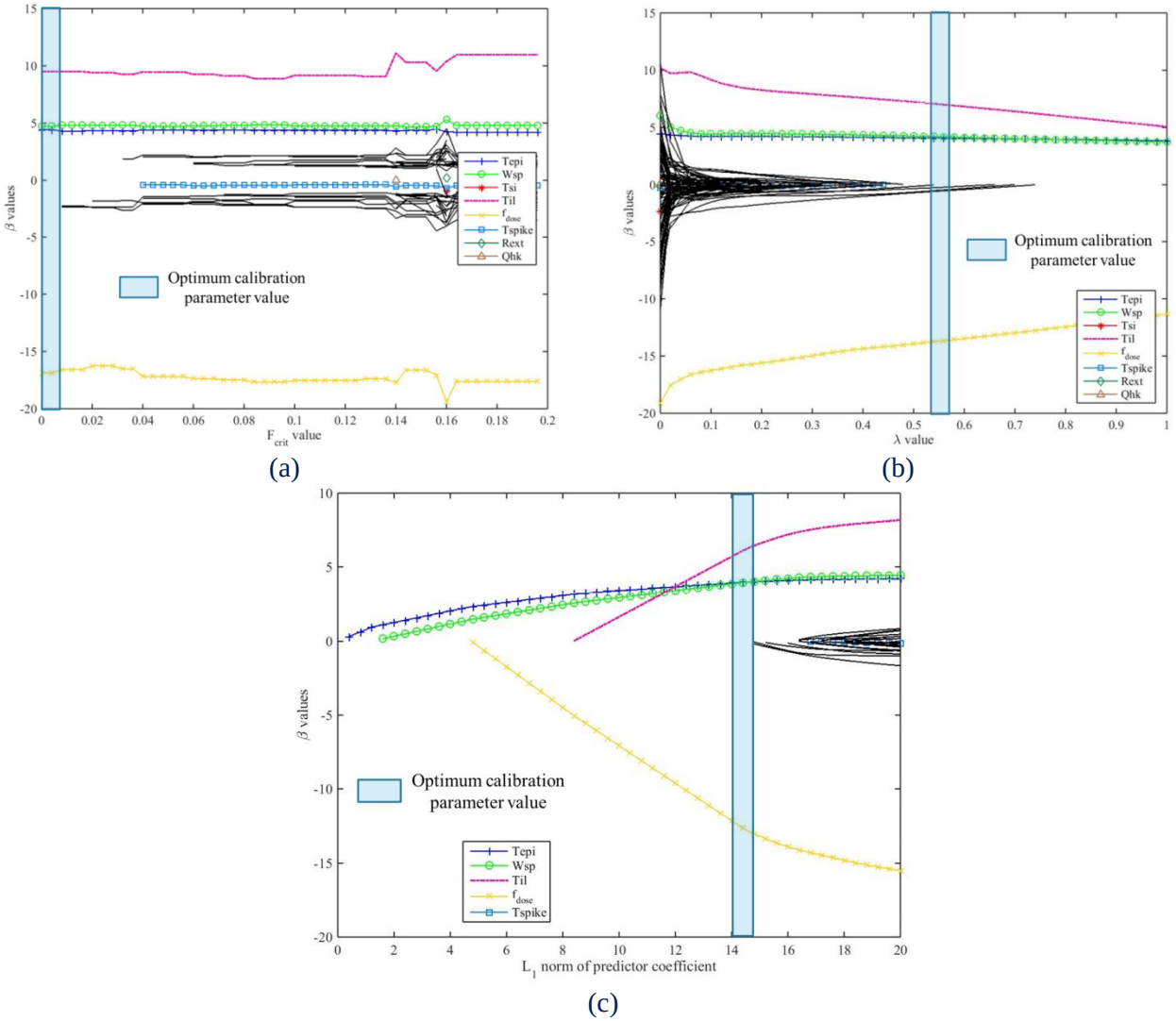


Figure 5-26: σ PCM constructed using stepwise regression (a), LASSO (b) and LARS (c) depending on their respective calibration parameter.

Figure 5-27 presents the results of $\mu_0.C_{ox}$ PCM construction. Stepwise regression and LARS method only includes T_{il} predictor, that is by far the most influent parameter on $\mu_0.C_{ox}$ as shown in Figure 5-19 (c). LASSO mistakenly includes Tepi. K-Fold CV and LOOCV indicate a large optimal area, such that Tsi can be justifiably included in the model whereas stepwise regression and LARS considered the effect of Tsi negligible.

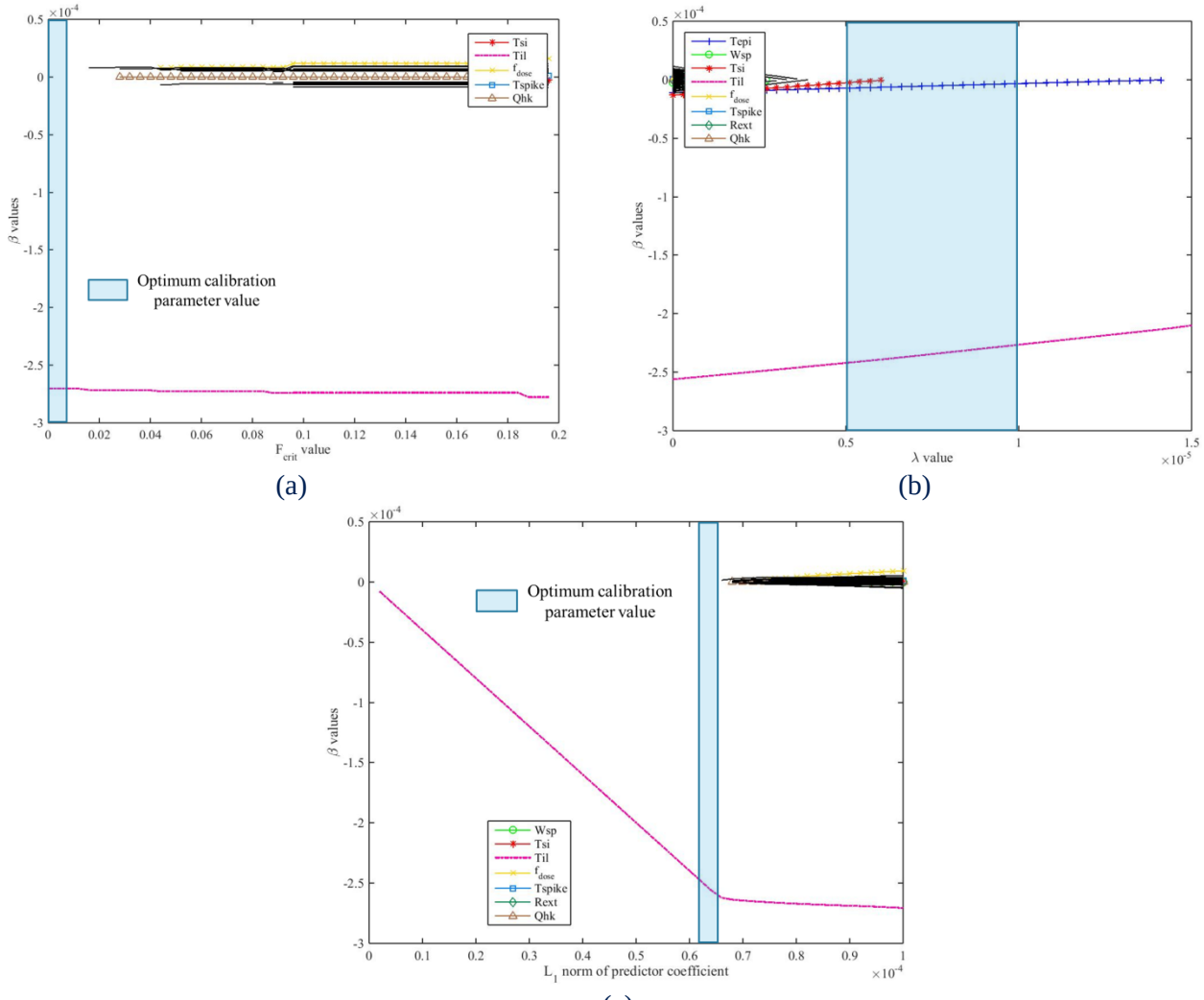


Figure 5-27: $\mu_0.C_{ox}$ PCM constructed using stepwise regression (a), LASSO (b) and LARS (c) depending on their respective calibration parameter.

Figure 5-28 shows the result of θ_2 PCM building. It is very similar to $\mu_0.C_{ox}$ as suggested by the extraction of model parameters on the TCAD simulated DOE in Figure 5-19 (c) and (d). PCM only includes T_{il} parameter. Compared to $\mu_0.C_{ox}$, θ_2 extracted on TCAD simulated DOE strongly depends on Qhk. This dependence is not accounted for by the PCM here. This is because of the small variance of Qhk. Indeed in these simulations, Qhk dispersion represents only 2% of Qhk variation simulated in the DOE.

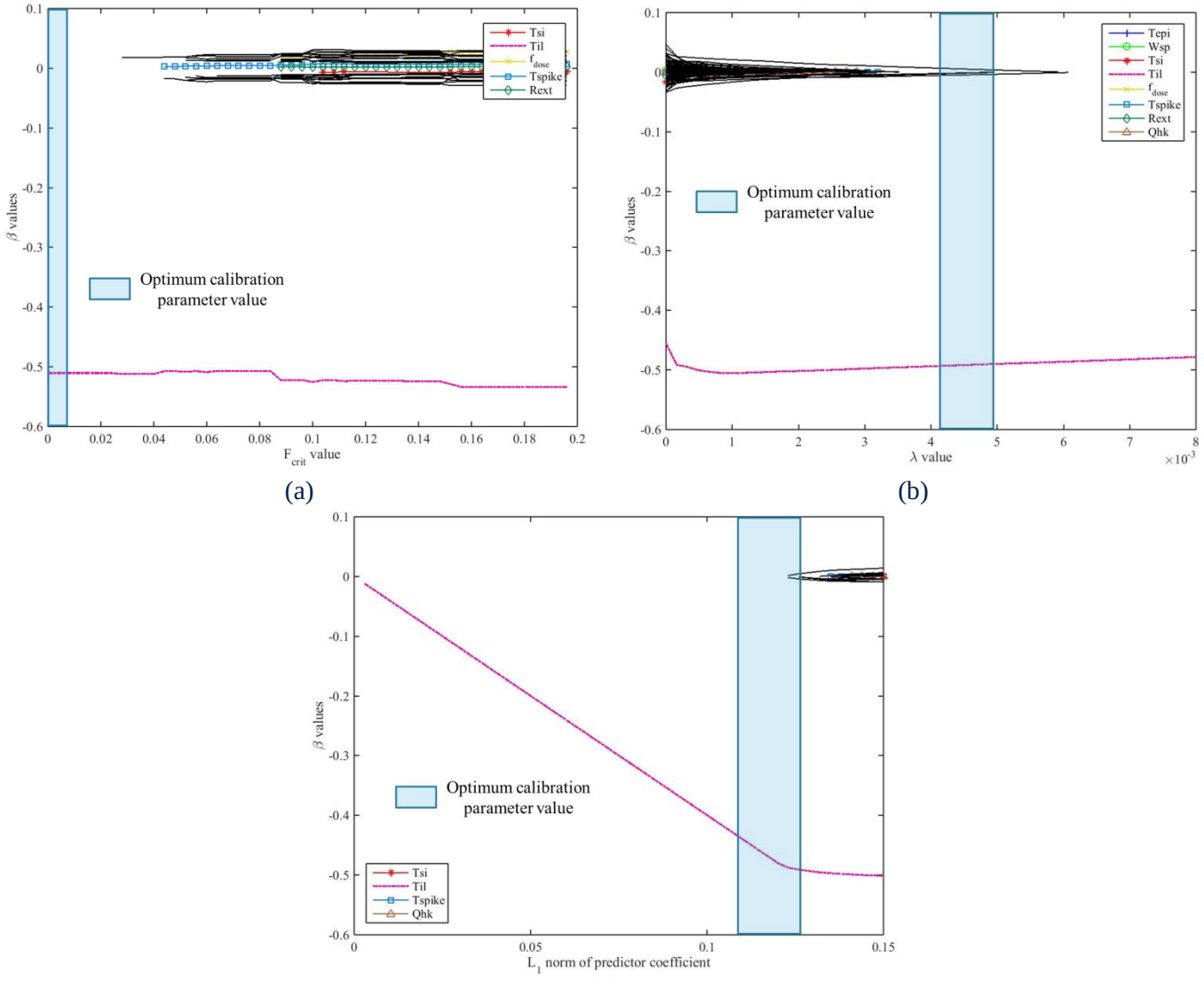


Figure 5-28: θ_2 PCM constructed using stepwise regression (a), LASSO (b) and LARS (c) depending on their respective calibration parameter.

V_{tlin} PCM construction results for short channel device are shown in Figure 5-29. We have seen that V_{tlin} is strongly dependent on T_{il} and Qhk. In addition, short channel V_{tlin} also slightly depends on Wsp. All these parameters are included in stepwise regression, LASSO and LARS PCMs. Tepi has been also included in the PCMs although its influence is limited regarding the results of TCAD simulated DOE. On the contrary, Tsi is not accounted in the model although its influence is not negligible as suggested by TCAD simulated DOE results. This is due to the very limited variance of Tsi in these simulations. LASSO and LARS may also mistakenly include fake predictors.

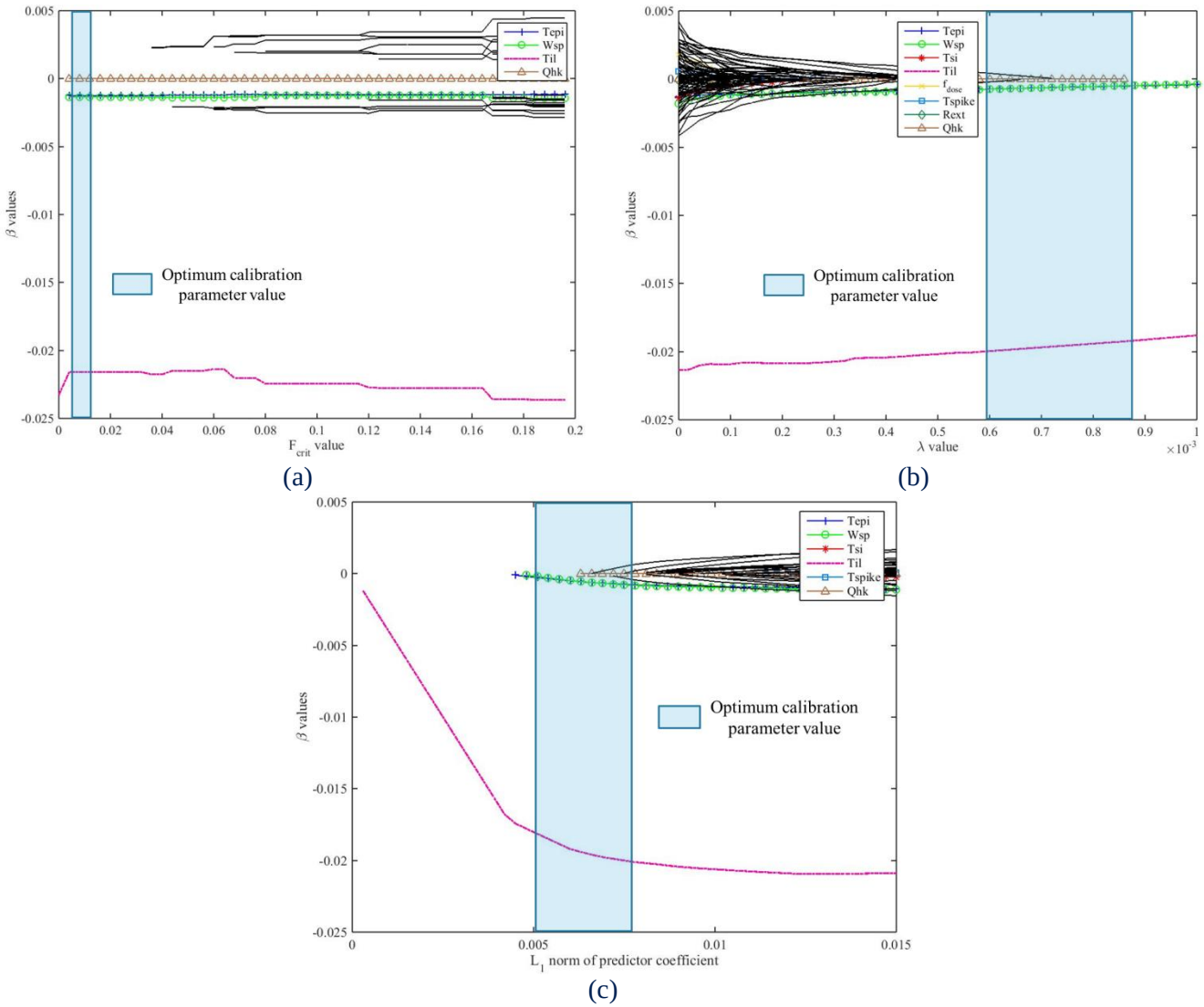


Figure 5-29: V_{tlin} PCM constructed using stepwise regression (a), LASSO (b) and LARS (c) depending on their respective calibration parameter.

Figure 5-30 shows the results of V_{tsat} PCM construction. V_{tsat} depends more or less on every process parameters as discussed previously. Constructed PCMs only includes Tepi, Wsp, Qhk, fdose and T_{il}. These are the most influent parameter as shown in Figure 5-19 (f). Again LASSO and LARS mistakenly includes fake predictors in the model with very low coefficients.

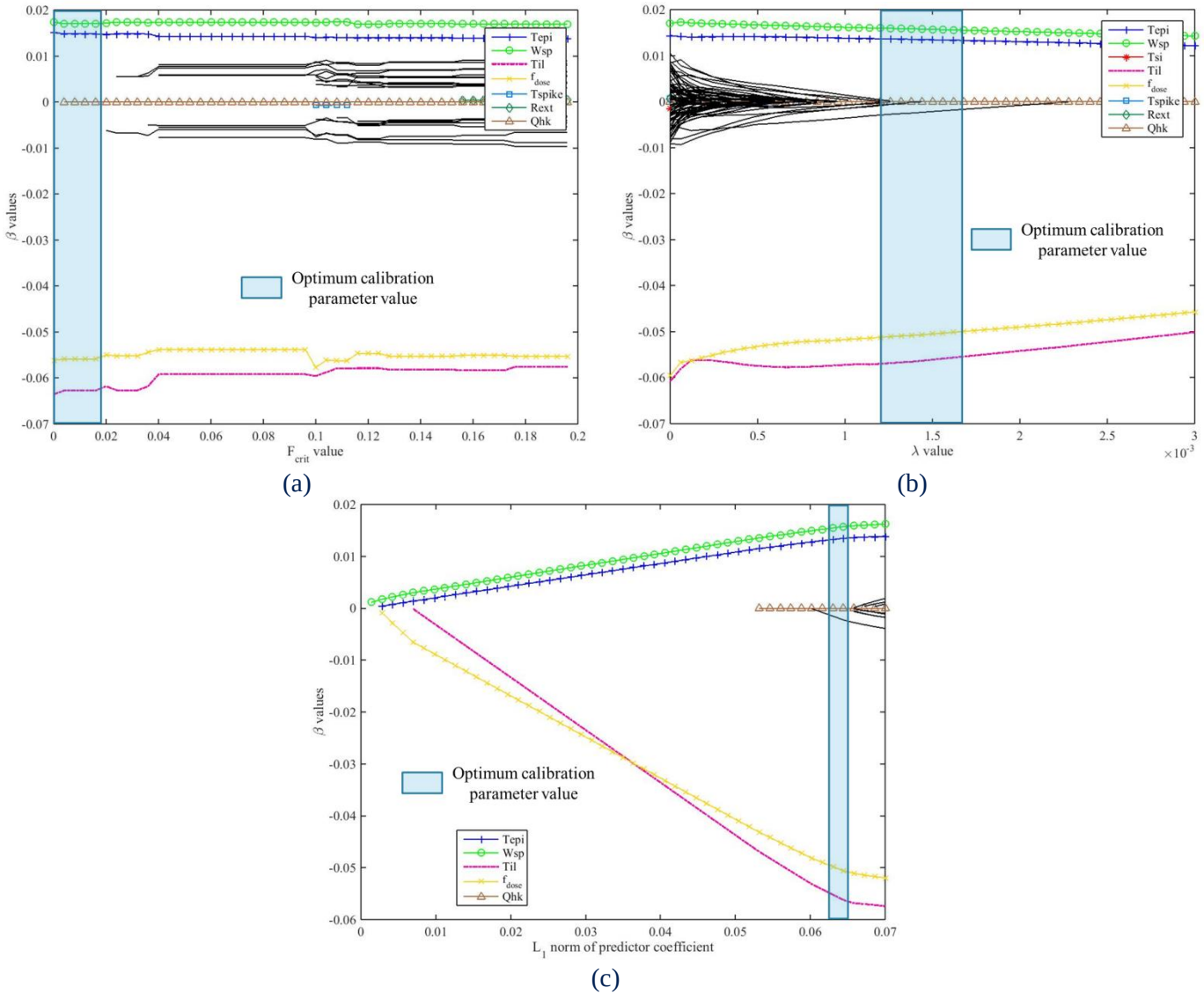


Figure 5-30: V_{tsat} PCM constructed using stepwise regression (a), LASSO (b) and LARS (c) depending on their respective calibration parameter.

Similarly to V_{tsat}, $v^*.C_{ox}$ depends on every process parameters as shown in Figure 5-19 (g). In these simulations, only fdose, Tepi Wsp and T_{il} have been included. These parameters are the most influent one. Although its influence is significant, Qhk is not present in the models because its dispersion is limited in these simulations. LASSO includes again fake predictors but with very low coefficients.

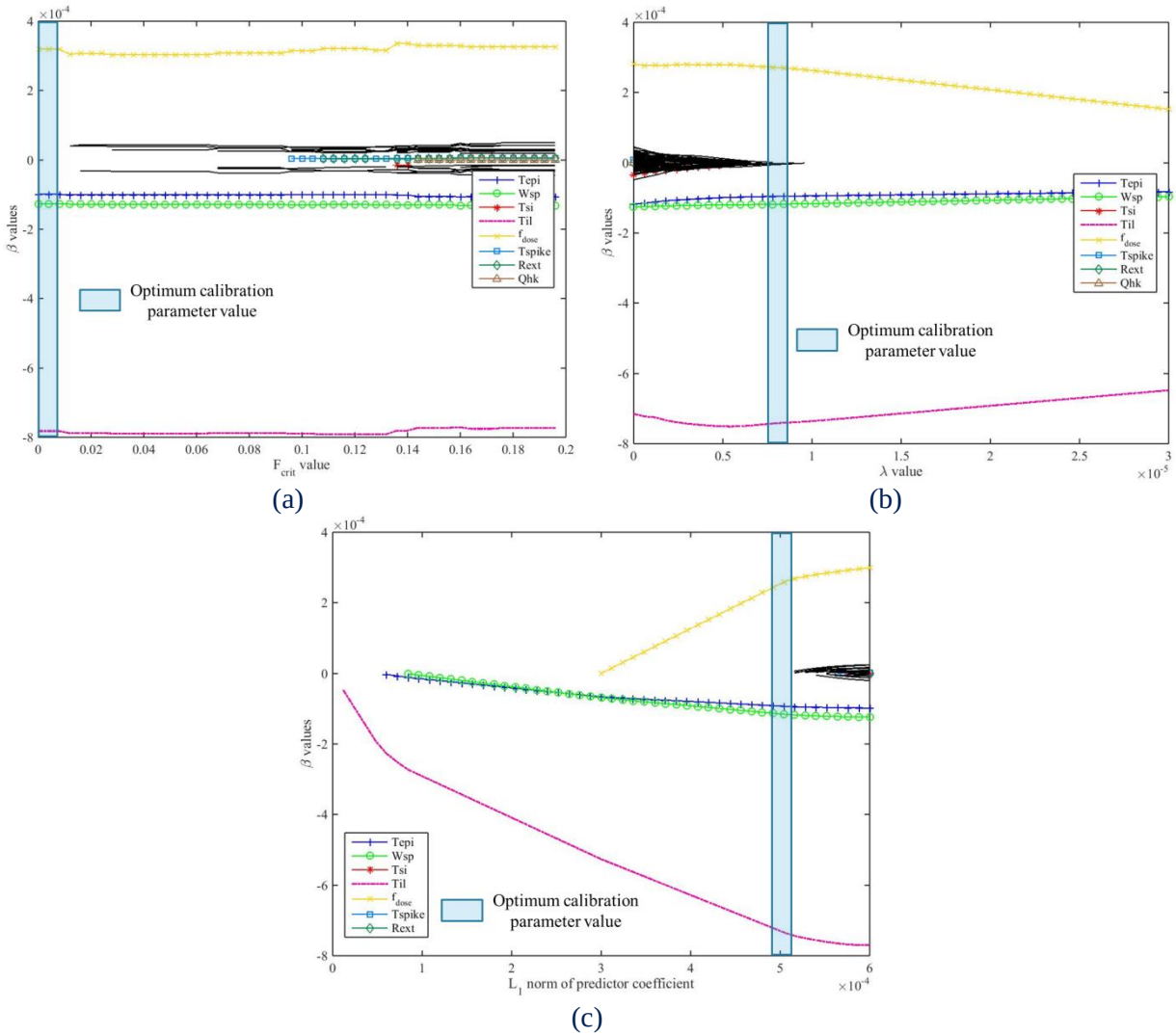


Figure 5-31: $v^*.C_{ox}$ PCM constructed using stepwise regression (a), LASSO (b) and LARS (c) depending on their respective calibration parameter.

To conclude this studie about PCM construction using data gathered over a single wafer, we have seen that using stepwise regression, LASSO and LARS algorithm, it is in most cases possible to distinguish process parameters that actually influence the considered model parameters from those who don't. LASSO is more prone to mistakenly include fake process parameters compared to stepwise regression. However LASSO has the advantage to yield continuous results depending on λ parameter. Consequently, predictors included in the model can be ranked from the most to the least likely related to observation variance. Considering this ranking and comparing the results of the three methods it is easy to find out if a predictor has been mistakenly introduced to the PCM. This likelihood ranking is not possible using stepwise regression. This is why it is recommended to compare the results of the three methods.

Compared to OLS, variable selection methods are much more efficient in order to build PCM. Indeed, variable selection successfully removed fake predictors whereas OLS included all of them in the model.

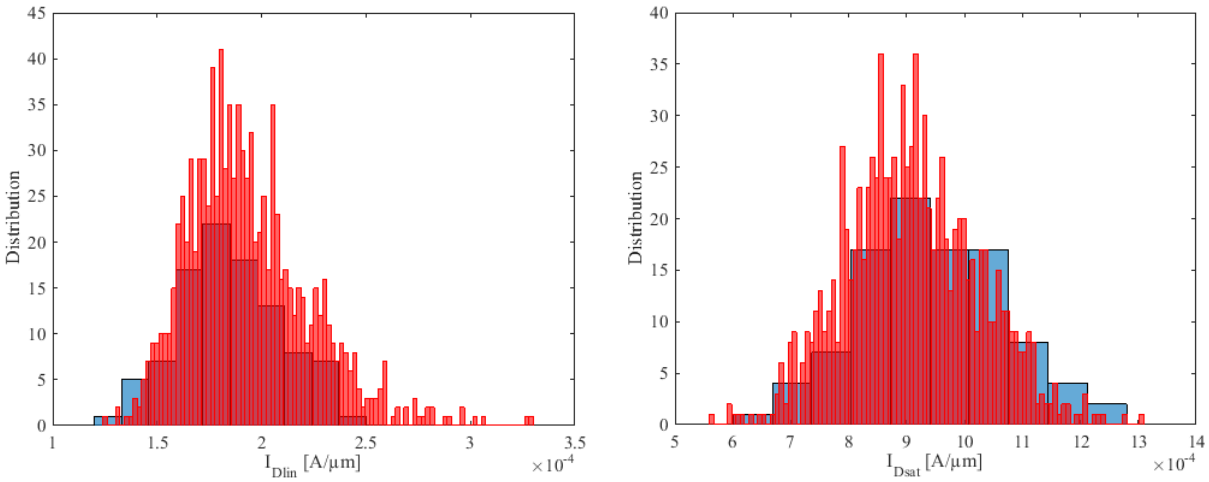
It should be noted that even if the variable selection works properly, PCM that results from observation at die scale over a wafer with small variation of process parameters across the wafer are different from those obtained from DOE observations. Indeed, the observation range being different, some process parameters have no significant impact on model parameters and are omitted in the PCM.

In order to be able to construct PCM that includes all relevant process parameters, PCM construction should be carried on using a DOE where each experiment are process on a wafer. Each wafer should then be measured at die scale. A full DOE processed with one wafer per experience would provide sufficient process parameter variance and die scale observations would limit the uncertainty about observations.

5.5 Using PCM to model and optimize within-wafer variability

In previous paragraph, we have seen that our PCM construction procedure is able to model process and electrical parameters relationships at wafer level. In this paragraph we will use PCM previously obtained in §5.4.3.2 to model within-wafer variability. We will show its accuracy against TCAD simulations and then exploit it in order to investigate the process origin of within-wafer variability and give guidelines for variability optimization.

For the sake of demonstration, we have propagated process parameters variability using Monte Carlo draws and the constructed PCM in §5.4.3.2. The results are compared with drain current dispersion obtained by TCAD simulations following the simulation setup developed in §5.4.1. In order to propagate process parameters dispersion and model electrical variability we have used 1000 Monte-Carlo draws. Considered process parameter dispersions are the one used in the simulation setup (see Table 5-3). Comparison between predicted and simulated I_{Dlin} and I_{Dsat} within-wafer variability is shown in Figure 5-32. A proper agreement is obtained, showing that the constructed PCM is able to model within-wafer variability.



(a) (b)

Figure 5-32: Within-wafer distribution of I_{Dlin} and I_{Dsat} modeled and simulated using TCAD

Using this variability model, it is then possible to diagnose its process origins and optimize it. Indeed, following propagation of variance developed in Chapter 4, if the model can be locally linearized in the model parameters dispersion range, within-wafer variability yields:

$$\sigma_e^2 = J^T \cdot \sigma_m^2 \cdot J \tag{156}$$

$$J_{i,j} = \frac{de_j}{dm_i} \tag{157}$$

where e and m denote electrical and model parameters respectively. Index i and j go through the number of considered model and electrical parameters respectively. σ_e^2 and σ_m^2 stand for the

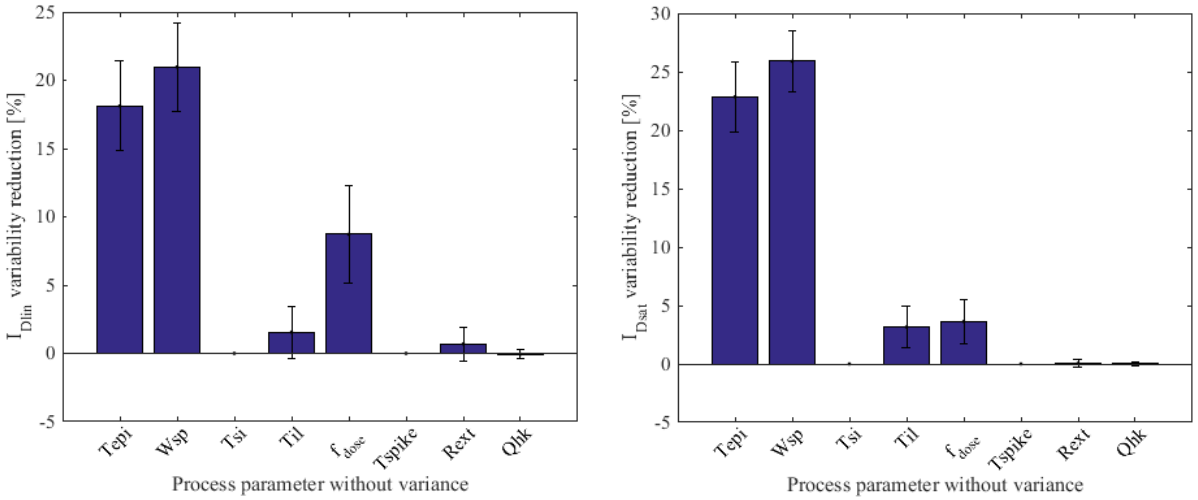
covariance matrix of electrical and model parameters respectively. J is the sensitivity matrix with respect to model parameters. This expression can be reformulated following:

$$\sigma_{e_j}^2 = \sum_i \sum_k \frac{de_j}{dm_i} \cdot \mathrm{Cov}(\mathrm{m_i}, \mathrm{m_k}) \cdot \frac{de_j}{dm_k} \tag{158}$$

Since model parameters are now modeled by their respective PCM, $\sigma_{e_j}^2$ can be further developed, replacing model parameters in (158) by process parameters:

$$\sigma_{e_j}^2 = \sum_i \sum_k \frac{de_j}{dp_i} \cdot \mathrm{Cov}(\mathrm{p_i}, \mathrm{p_k}) \cdot \frac{de_j}{dp_k} \tag{159}$$

In this expression, we can see that the total within-wafer variability depends on drain current sensitivity to process parameters and on the process parameters dispersions. Thus optimizing $\sigma_{e_j}^2$ consists in either reducing process parameters variance or reducing $\frac{de_j}{dp_i}$. Reducing process parameters variance relies on process tool optimization. Consequently we need to rank process parameters according to their contribution to drain current variability, so that we can draw the best benefits with a minimum effort in tool optimization. In order to rank process parameters we have run a statistical test. The test consists in estimating the drain current variability with 1000 Monte Carlo shots while setting one process parameter variance to zero at a time. The test is repeated for each process parameter. Estimated drain current variability is then compared to the actual one (where all process parameters are variable). Figure 5-33 shows the reduction in drain current variability depending on the process parameter whose variability has been suppressed. We see that a large gain in variability can be obtained by reducing Tepi and Wsp variability for both I_{Dlin} and I_{Dsat}. On the contrary it is useless to work on Tsi, Tspike, Qhk or contact resistance (Rext) variability since their contributions are small.



(a) (b)

Figure 5-33: Expected drain current variability reduction by suppressing the variability of one process parameter at a time.

A warning should be put while considering this plot. One should not deduce that drain current does not depend on Tsi, Tspike, Qhk or Rext. Actually it does but the considered variability for these parameters is so small that it does not impact significantly the drain current variability.

However, studying and optimizing drain current sensitivity to process parameters can help reducing drain current variability as well. Indeed, this sensitivity is modeled by the term $\frac{de_j}{dp_i}$ in (132). Reducing

it relies on two options. Since it depends on the model and physics underlying the device, the first option is to investigate other technologies, with drain current less sensitive to process parameters. However it is a tedious option. The other option is to shift the device operating point by shifting process parameters values. Indeed $\frac{de_j}{dp_i}$ depends on considered average p values (the operating point). However, shifting process parameters values will change the average drain current value as well (and maybe not in a good way). In order to reconcile variability and performance targets, a two-objective optimization task should be carried out. To do so, the objectives function is the electrical parameter and its variability. Then SPC can be used in order to define a penalty function in order to run optimization algorithm [177].

5.6 Effect of local random variability and measurement noise

As it has been discussed in §5.4.3.2, so far we, have built PMC at the wafer level. In order to get a PCM that includes all relevant process parameters with significant dispersion, we suggested building PCM using a full DOE processed on silicon. Each experiment would be run on one wafer. Each wafer would be monitored at die scale in order to reduce uncertainties about process and electrical parameters. In this paragraph we investigate this application using synthetic data. The critical point that will limit the efficiency of the method, while applying it on silicon measurements, is the residual noise or uncertainty in the observations. Indeed, electrical measurements are impacted by local variability and noise. This uncertainty will first propagate through the model extraction procedure and then bias the PCM construction results. This noise should be low enough to ensure robust results. Here, we investigate this question using synthetic data where local variability, within-wafer variability and noise are simulated.

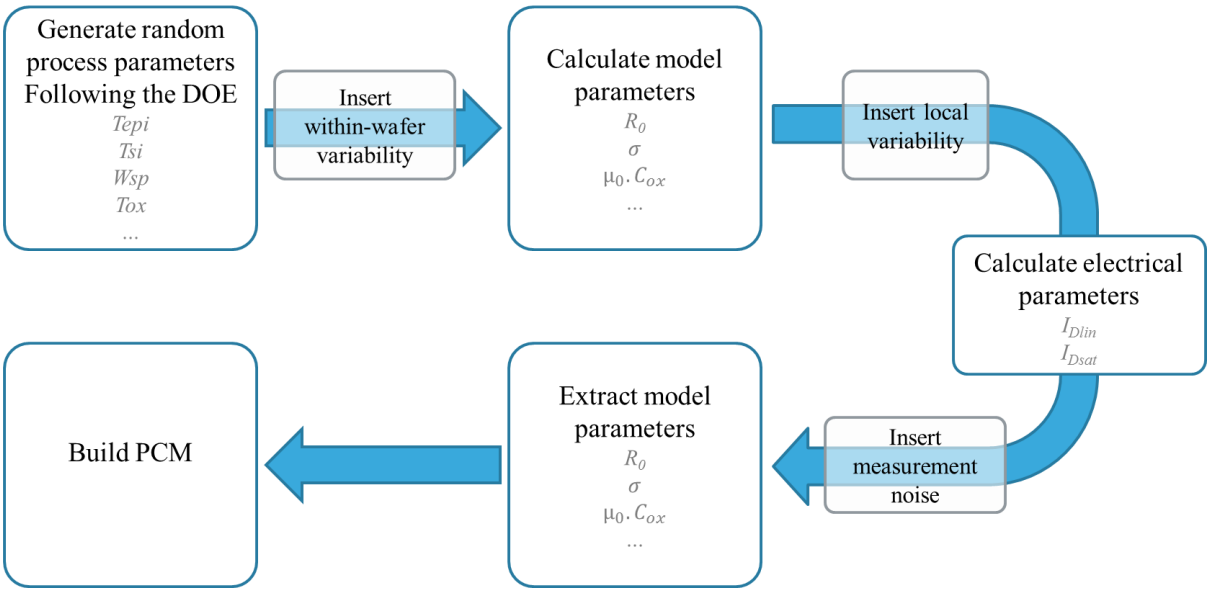


Figure 5-34: Test flow used to check PCM construction procedure robustness to noise and variability

The test method is described in Figure 5-34 as a flow chart. First it consists in generating synthetic data for process parameters. The dataset is constructed such that it mimics the kind of dataset that would be measured if experiments are run on silicon. In other words, process parameters values are generated following a specified DOE. Each experiment is run using one wafer and each wafer is composed of many sites. Then within-wafer variability is simulated by adding a random signal to these process parameters. This signal is centered on each experiment value and dispersed over dies of the corresponding wafer. Within-wafer dispersion is set at 2% of the averaged values of process

parameters. Since process parameters are known at die scale, within-wafer variability does not contribute as a source of uncertainty.

Then, considering extracted PCM in §5.4.2, model parameters are calculated for each experiment. Local variability is simulated adding a random signal to these model parameters, according to values found in literature. Using the drain current model, I_{Dlin} and I_{Dsat} are then calculated. In order to simulate measurement noise, a random signal is added to I_{Dlin} and I_{Dsat} values according to predefined noise threshold.

The second step consists in going backward using synthetized I_{Dlin} and I_{Dsat} to find the PCM used initially to create synthetic data. Using the extraction procedure, I_{Dlin} and I_{Dsat} model parameters are extracted for each experiment. Then PCM are built on these extracted model parameters. In order to test if variable selection is efficient, extra process parameters are added to the PCM building method input. These extra parameters are randomly drawn, independently to the other process parameters. If variable selection method is sufficiently robust, it will rule these extra parameters out of the PCM.

In order to accurately reproduce the experimental conditions, the star points of a central composite DOE are considered for process parameters effect investigation. This design considers 3 levels for each process parameter and comprises 17 experiments. Noting the different level -1, 0 and 1, this DOE is represented in Table 5-4. Correspondence between the level and process parameters values are shown in Table 5-5.

Experiment\Parameter	Tepi	Wsp	Tsi	T_{il}	Fdose	Tspike	Rext	Qhk
1(reference)	0	0	0	0	0	0	0	0
2	-1	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0
4	0	-1	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0
6	0	0	-1	0	0	0	0	0
7	0	0	1	0	0	0	0	0
8	0	0	0	-1	0	0	0	0
9	0	0	0	1	0	0	0	0
10	0	0	0	0	-1	0	0	0
11	0	0	0	0	1	0	0	0
12	0	0	0	0	0	-1	0	0
13	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	-1	0
15	0	0	0	0	0	0	1	0
16	0	0	0	0	0	0	0	-1
17	0	0	0	0	0	0	0	1

Table 5-4: Face centered composite DOE with process parameters.

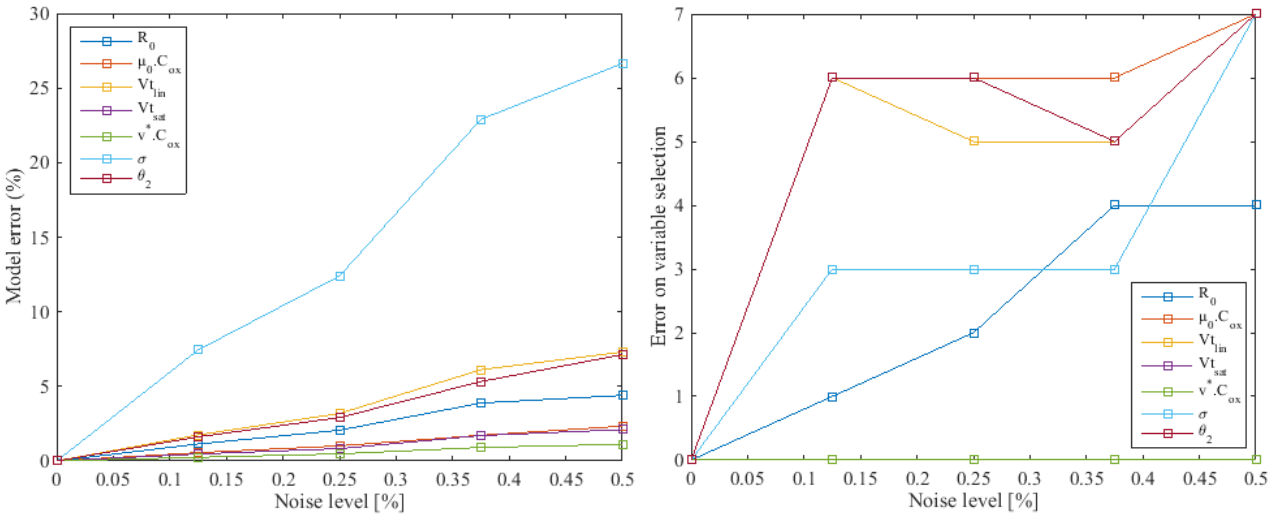
Variable Level	Tepi	Wsp	Tsi	Til	fdose	Tspike	Qhk	Rext
-1	12	8	5	0.8	0.5	900	10^{10}	70
0	14	10	6.5	1	1	1000	10^{12}	100
1	16	12	8	2	1.5	1100	10^{13}	130

Table 5-5: Process parameters values used in the DOE depending on the level.

Considering that each experiment is run using one wafer, a lot would be sufficient to run this DOE. Each wafer contains many sites. In order to simulate their impact, we consider that 9 of them will be

monitored. Each of these dies is probed independently, so process and electrical parameters are known at die scale. Global variability is simulated with a normally distributed random signal with $\sigma = 2\%$ applied on process parameters. Then local variability contribution is accounted through model parameters. Local variability mostly impact Vt, β and Rsd variability. Considering devices studied here, an average value of $\sigma_{\Delta V_{TH}}$ is $\frac{2}{\sqrt{WL}}$ mV.µm, $\sigma_{\frac{\Delta\beta}{\beta}} \cong \frac{1\%}{\sqrt{WL}}$ µm and $\sigma_{\frac{\Delta R_0}{R_0}} \cong 20\%$. These values have been found in recent publications [178]. Finally, measurement noise is added through I_{Dlin} and I_{Dsat} values. Noise in electrical measurements depends on the integration time used for measurements. Typical measurement time for in line parametric test is about few milliseconds. Associated noise does not exceed 1%.

For each site we have one dataset of I_{Dlin} and I_{Dsat} values. Each of them is composed of drain currents synthetized at Vg = [0.7, 1, 1.1] V. 11 gate lengths have been considered for the extraction. The first test has been run considering no local variability and a noise level ranging from 0 up to 0.5%. Error made on constructed PCM is shown in Figure 5-35.



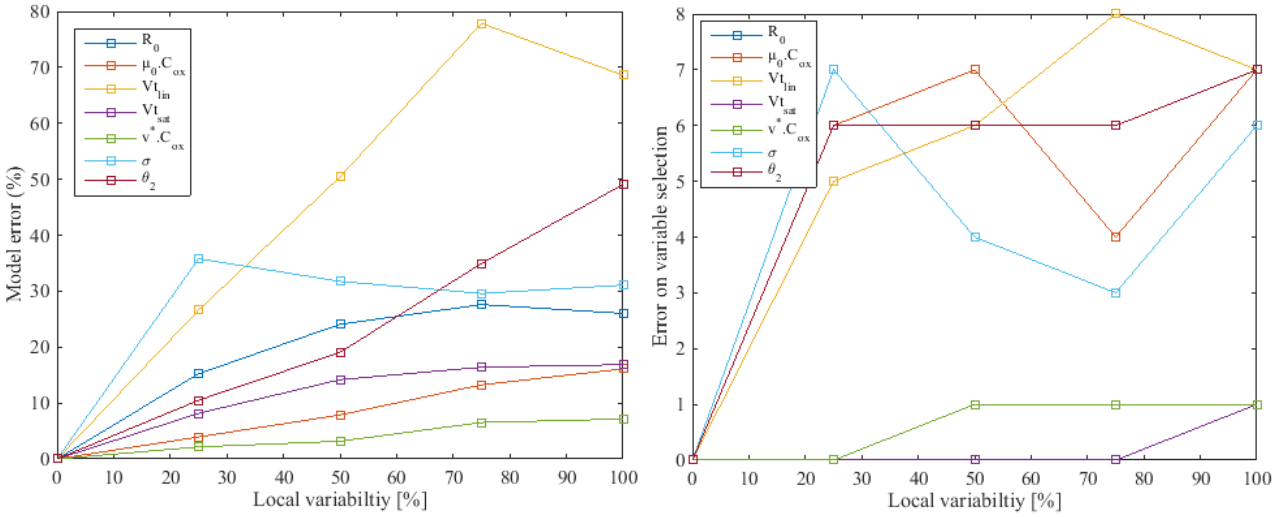
(a) (b)

Figure 5-35: RMS error made on drain current (a) build from synthetic data and error on selected variables (b).

Figure 5-35 (a) shows the RMS error made on model parameters depending on the simulated noise level. Figure 5-35 (b) shows the number of variables mistakenly added or omitted in the PCM. We see that PCM for saturation regime, model parameters (v^*C_{ox}, and Vt_{sat}) are fairly insensitive to noise. Variable selection works perfectly and error on PCM coefficient is rather low. However a small amount of noise already compromises PCM construction for linear model parameters. Error on variable selection stems from the inclusion of "fake" predictors or omission or relevant predictor, if those have a negligible contribution to the PCM.

There are two solutions in order to reduce the impact of noise in measurements: i) increasing the integration time during measurement, ii) increasing the number of measurement points (more L and/or more V_G).

The second test has been run considering no measurement noise but local variability going from 0 up to 100% of noise reported in recent publication [**178**]. Error made on constructed model parameters is shown in Figure 5-36.



(a) (b)

Figure 5-36: RMS error made on PCM coefficient (a) build from synthetic data and error on selected variables (b).

Impact of local variability on PCM RMS error is much stronger than the effect of noise in measurements but variable selection performance is similar. This is due to the fact that "fake" predictor included in the PCM have much stronger coefficients and induce larger error. Again, saturation parameters are less sensitive than linear ones. In order to reduce the impact of local variability, the solution consists in measuring arrays of transistors instead of isolated transistors. Following this procedure, local variability is averaged over the whole array's area rather that a single transistor area. Generally speaking, local variability is proportional to $1/\sqrt{W \cdot L}$. This means that using a 20x20 array of transistors, the impact of local variability would be divided by 20. Moreover, measuring multiple transistor in the same time reduce the measurement noise significantly. Indeed, considering 400 transistors instead of one, the equivalent number of integration points for one measurement is multiplied by 400. Following the Standard Error (SE) of the mean formula (160), the noise level is inversely proportional to the square root of the measurement sample size (for a fixed integration time).

$$SE_{mean} = \frac{s}{\sqrt{n}} \tag{160}$$

In (160), s and n are the standard deviation and the size of the sample. Thus accounting for 400 transistors would divide the noise by 20.

Thus using array of transistor and increased number of measurements and integration time can significantly reduce the local variability and noise burden.

To investigate this option we have run the same test as before considering this time an array of 20x20 transistors. Measurement noise and local variability have been both accounted for, dividing their amplitude by 20. Results are shown in Figure 5-37. Performances are much better, reducing the maximum PCM error below 2%. Variable selection is improved but fake variable are still included in the PCM. However here, their coefficients are very low, limiting the PCM error.

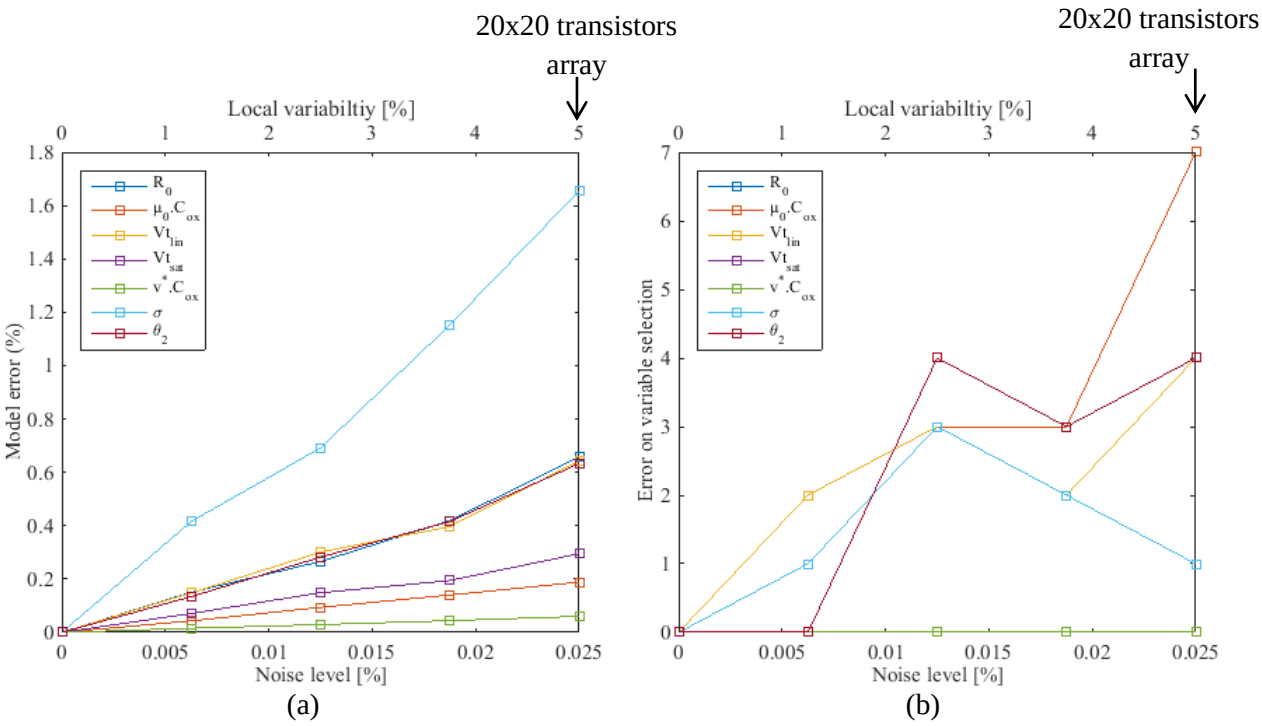


Figure 5-37: RMS error made on PCM (a) build from synthetic data and error on selected variables (b).

5.7 Conclusion

In this chapter, we have introduced the concept and benefits of PCM. Different approaches to build PCM have been investigated, namely stepwise regression, LASSO and LARS, along with method to test the predictability of the results and to calibrate the methods. These methods are K-fold CV, LOOCV and bootstrap. Stepwise regression, LASSO and LARS combined with K-fold CV, LOOCV and bootstrap methods have been tested against synthetic data. Their ability to perform variable selection and to find back the proper polynomial formula used to generate synthetic data has been tested against artificial noise with ill-posed problems. It has shown to be able to handle 5% of noise with only 25 observations against 50 predictors.

Then we have applied stepwise regression, LASSO and LARS to build PCM based on simulation results. First we have seen that building a polynomial formula that directly link electrical parameters (I_{Dlin} and I_{Dsat}) with process parameters is not appropriate since it would require building one PCM per V_G and gate length, making the model cumbersome. Moreover the model quality is not satisfying.

As an alternative, PCM have been built for model parameters (R_0, σ, $\mu_0.C_{ox}$, θ_2, Vt_{lin}, Vt_{sat}, $v^*.C_{ox}$). Using this approach we only need one PCM per model parameter to be able to model linear and saturation drain current over the entire strong inversion regime. In addition to this advantage, PCMs have shown to be more accurate for these parameters. Indeed, these model parameters are more elementary compared with I_{Dlin} and I_{Dsat} and their dependence with process parameters are thus simpler. Using OLS, we have then shown that linear polynomial formulas are suited for those PCM except for T_{il} and Tspike that have nonlinear relationships with model parameters. This problem can be solved either by considering a reduced range for process parameters or by using empirical nonlinear formula along with variable shift for the model. This has been exemplified modeling σ and Tspike relationship with an exponential formula.

However, OLS does not perform variable selection and obtained PCM included unphysical dependences. Moreover, it has been pointed out that dealing with silicon measurements induces two

extra issues. Frist, we have to deal with within-wafer variability and, second, there are a large number of uncontrolled process parameters. In order to improve the accuracy, process and electrical parameters should be monitored at the die level. However, this lead to deal with randomly distributed process parameters and ill-posed problems (since there can be more process parameters than observations) instead of a regular DOE. In order to test the applicability of PCM construction method on silicon measurements, we have simulated experiences with randomly distributed process parameters. Using OLS to build PCM has shown to be a poor solution since "fake" predictors are included in the PCM and those could have non negligible coefficients. Alternatively, using stepwise regression, LASSO and LARS algorithm, we have seen that it is possible to obtain good results even in the presence of randomly drawn "fake" predictors. Thus the method is efficient in selecting relevant process parameters to build PCM even if the problem is ill-posed. However we have seen that some process parameters are not sufficiently dispersed to have a significant impact and are not included in the final PCM. Thus we suggest using both DOE and die scale observation in order to build proper PCM.

In order to assess the impact of local variability and measurement noise on PCM construction robustness, tests have been run using synthetic data. Starting with a DOE that includes different process parameters, model parameters and drain currents have been synthetized using the drain current model and PCMs built on TCAD results. Local variability and measurement noise have been simulated through a random signal added to the synthetic model parameters and drain current values. These synthetized data drain currents have then been used as input to the extraction method. Extracted model parameters have in turn been used to build PCMs. The error between PCM used as input and the one extracted revealed the impact of local variability and measurement noise. The PCM construction is compromised by the impact of noise and local variability considering a single transistor measured with short measurement time. However, we have shown that this problem can easily be bypassed using array of transistors. We recommend using 20x20 transistors array, without increasing the measurement time in order to reach a proper noise level. Of course increasing the measurement time would increase the PCM construction robustness as well.

As a perspective I shall note that, although different methods to build PCM and to calibrate these methods have been investigated, this study is not comprehensive. Indeed, domain of variable selection and machine learning are highly active research area and many alternatives (probably more efficient one) can be found in literature. Methods developed here have been chosen because they are widely used, well documented and reliable. For information purpose only, other interesting work related to this area can be mentioned. In term of variable selection, one can found random forests method introduced by L. Brieman **[179]**. This method, as well as those investigated in this work, only deals with linear polynomial models. In case one wants to test nonlinear model, other methods are more appropriate **[180]-[186]**. In term of PCM, we have investigated the use of linear polynomial formula. However, more flexible and efficient alternatives can be mentioned like, Kernel Nearest Neighbor (KNN) **[185]**, Feed-Forward Neural Network **[186][187]**, Suport Vector Machine (SVM) **[188]**. Considering PCM evaluation and calibration of PCM construction method, we extensively used cross-validation methods. Other approaches are available like S_p criterion, Akaike Information Criterion (AIC), Final Prediction Criterion (FPE), C_p criterion of Mallows or the small-sample corrected version of AIC. A more recent publication **[189]** briefly reviews these methods and proposes its own method inspired by Beran and Dümbgen **[190]-[192]**.

Chapter 6 :
Conclusion

Technological node succession has been slowing down recently [193] due to new technological challenges involved. Among these barriers we find an increased impact of process and local random variability due to the increased complexity of fabrication process and dimension scaling, in addition to the difficulty to reduce the channel length. Some of these challenges require the adoption of new architectures that are very different from the traditional one (bulk transistors). However these new architectures involve more efforts in order to be industrialized. Increase of complexity and development time imply larger financial investments. Even though the semiconductor market is large and its sales rising continuously as shown in Figure 6-1, the industry growth display less convincing trends, strongly dependent on global economic climate [194].

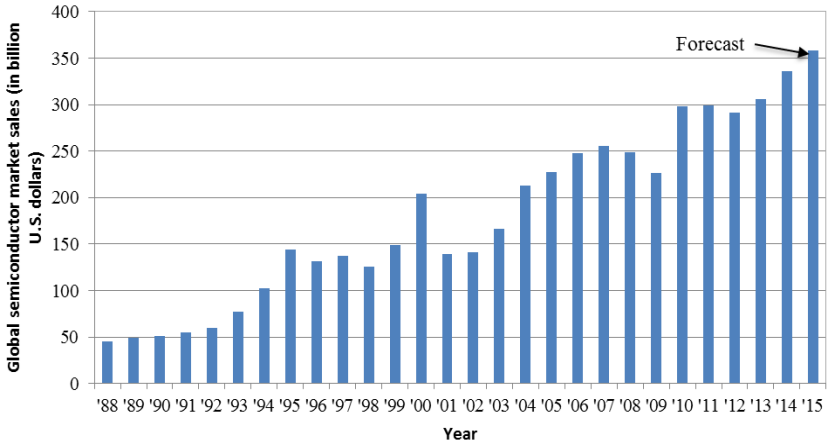


Figure 6-1: International semiconductor market sales per year from 1988 up to 2015 [195][196][197].

Consequently, investment margin shortens and R&D sector (that requires large funding with long term payback period) will have to deal with this chance. A recent survey asked semiconductor executives around the world what they see as the biggest issues facing the semiconductor industry during the next three years. The first answer is the increase in R&D costs followed by technology breakthroughs and high cost for plant and equipment [198]. Thus there is a real need for improving development and optimization of device fabrication. This work gives some leads in order to meet these expectations. In order to develop and optimize new transistors, engineers heavily rely on successive trials and expert knowledge. This approach appeared to be the most efficient and reliable until now. However, with the increasing complexity of new architectures and higher variability, this approach tends to require more and more trials, increasing dramatically the development cost. In order to address this issue, the idea is to minimize the number of trial in order to find the optimal fabrication process. The optimal process is the one that would lead to a device whose electrical performances and dispersion match predefined targets.

A way to find this optimal process without heavily relying on silicon processing is to use models and TCAD simulations. Indeed an accurate and fast-to-compute model that maps the relationships between process and electrical parameters can be used as an input to an optimization algorithm which in turn can find the optimal process. TCAD is a physically based and reliable tool. However each simulation is time consuming (in the order of magnitude of an hour). Since most of the optimization algorithms require a large number of evaluations (more than 10^3), relying only on TCAD would never yield timely answers. Moreover TCAD calibration is a tedious task, requiring large physical investigations. On the contrary, compact model (such as BSIM, PSP, …) are fast to compute and require only a full electrical characterization of the device to be calibrated. However most of model parameters are not directly linked to process parameters, per say, some parameters have a complex physical interpretation. For example, the physical interpretation of effective channel length (that is widely used

in compact model) has been the subject of many studies **[50][83][91][99]-[101][106][114][127][130]**. Thus, using compact model to optimize electrical performances cannot yield any accurate process flow.

The idea that has been developed in this thesis is to combine both TCAD and compact model in order to build and calibrate what is called Process Compact Models (PCM). PCM is an analytical model that maps the relationships between MOSFET's process and electrical parameters. It draw the benefits of both TCAD (since it relates electrical and process parameters) and compact model (since it is analytic and thus fast-to-compute). Our PCM is decomposed in two stages. Starting from process parameters, the first stage is formed of multiple polynomial formulas that relate process with the model parameters of a typical threshold voltage based compact model. The second stage is the compact model, yielding electrical parameters as output. An input/output scheme of this two-stage PCM is presented in Figure 6-2.

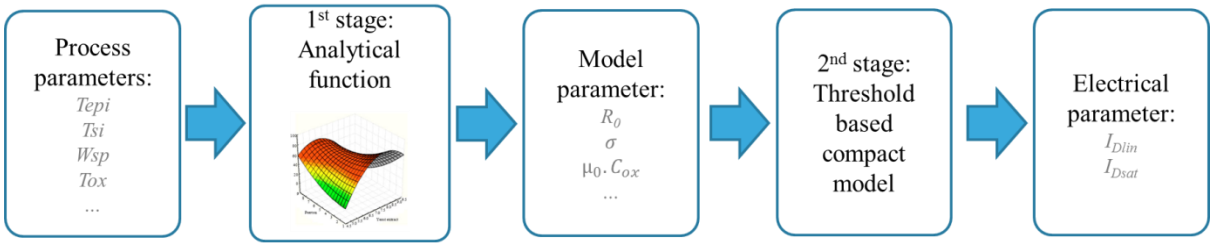


Figure 6-2: Scheme of the two-stage PCM

6.1 Summary of the thesis

After a short introduction, this manuscript details in chapter 2 the compact model used in the second stage that split linear and saturation drain current into model parameters such as access resistance, carrier mobility, threshold voltage… Derivation and physical background related to this compact model has been developed. Firstly, the MOS capacitance structure has been investigated to derive the inversion carrier concentration as well as the threshold voltage for the case of bulk devices. Then these equations have been adapted to the case of UTBB devices. The effects of channel doping concentration, ultra-thin channel and box on Vt and inversion charge density have been treated. A compact model for carrier mobility has been suggested, where surface roughness, remote Coulomb and phonon scattering as well as neutral defects, ballistic transport, saturation and injection velocity are accounted for. Then linear and saturation drain current formulations have been introduced, based upon proposed mobility, threshold voltage and inversion carrier concentration formulation. In real devices, compact models have to account for access resistance. Hence this aspect has been treated and analytical compact models for linear and saturation regimes have been adapted.

A large part of this thesis has been devoted to provide the right procedure to build and calibrate this PCM. It can be done using TCAD and/or silicon measurements. The procedure requires few drain current measured or simulated on transistors with different gate lengths, at different gate voltages. Using these data, model parameters are accessed using a specific extraction procedure developed in chapter 3. This method relies on few measurements for comprehensive monitoring of the production line. The method is decomposed into 3 steps. First step consists in extracting linear model parameters using linear least square fit. Then these values are used as a first guess for a nonlinear optimizer that refines parameters values. Finally, saturation model parameters are extracted using nonlinear least square fit.

This method has been tested to assess its robustness. Tests have been conducted on synthetized data against sample size and range. We have seen that data sample ranges and sizes available in silicon

measurements are too small to properly extract all model parameters. Removing successively each parameters from the model showed that θ_1, Vt_{LDR} and Lc are the least significant model parameters. Extraction test has been run once again considering cases where some of these parameters have been fixed. It showed that as soon as one parameter is removed, the extraction works fine. Thus removing one parameter allows robust extraction with a minimum bias in the model.

Following that study, the effect of measurement noise in the extraction procedure has been investigated. It revealed that a small amount of noise can lead to strong errors in model extraction. TCAD investigation of the mobility compact model showed that using both θ_1 and θ_2 in the model can lead to a high uncertainty about extraction results. Removing θ_1 allow more robust extractions against noise without making the parameter less meaningful. Noise tests have been conducted considering model parameters extracted on full I_D-V_G of nMOS and pMOS devices of 28 and 14 nm FD-SOI technologies and setting θ_1 to 0. Results showed reasonable level of noise in extracted model parameters considering 1% of noise in electrical parameters.

These test showed that attention must be paid to the model used for extraction. First we suggest setting $\theta_1$to 0 in order to reduce the impact of noise in measurements. Then, depending on the device, one or two parameters must be removed (Vt_{LDR} and/or Lc). In order to verify the validity of such simplifications, extraction results must be checked. Considering TCAD simulations, the physical coherence of the results has been checked against corresponding process variations. Considering silicon extraction, correlation plots have been performed, showing that model parameters are mostly uncorrelated. Uncorrelated parameters ensure the robustness of the extraction and enable drawing inferences of model parameter's variation impact on drain current.

The extraction procedure has been run on a TCAD simulated DOE. The DOE account for different process parameters (External resistance, epitaxial thickness, SOI thickness, spacer width, implanted dose, annealing temperature, insulting layer thickness, high-K thickness). We have shown that model parameters response to process variations is physically coherent, testifying on model parameters physical meaning and extraction robustness. Extractions have been run for nMOS and pMOS enabling a quantification of the impact of active dopant dose in the source-drain region as well as the junction profile on the drain current and model parameters.

Following the introduction of model parameter extraction procedure in chapter 3, we have applied it on silicon measurements in chapter 4 where 28 and 14 nm FD-SOI technologies have been investigated. It has been shown that model parameters variations depending on process variations are coherent and have been physically interpreted. A clear quantification of the impact of process variations has been enabled, showing that the method is efficient and robust while requiring only few measurements, making it suitable for industrial application.

Studying 28 nm FD-SOI using model parameter extraction enabled quantifying the impact of source drain implant dose and energy as well as DSA step. We have seen that extractions yield physically coherent results. Highly doped source-drain region resistance R_0 is lowered by higher implant dose and energy and by DSA. Both these process parameters directly influence the active dopant concentration. This means that highly doped source-drain region has remaining inactivated dopant before DSA. DSA activates them successfully. On the contrary LDR resistivity represented by σ is only dependent on implant dose and energy. Indeed DSA does not induce dopant migration and thus doesn't move the junction further toward the channel. Moreover this means that LDR dopants are already well activated before DSA and DSA has no activation effect in this region. However Vt_{LDR} extraction has evidenced that the junction position is sensitive to implant energy and dose. $\mu_0.C_{ox}$, θ_2 and Vt_{lin} have been shown to be constant, meaning that dopant does not penetrate into the metal gate

or channel. All these sensitivities can be quantified easily using this technique, bringing valuable information in terms of device optimization.

Studying 14 nm FD-SOI technology, it has been possible to evaluate the impact of HF cleaning time before epitaxy, carbon and phosphorous dose during in situ doped raised source-drain epitaxial as well as epitaxy thickness. Carbon has shown to increase R_0 by reducing dopant migration whereas increased phosphorous dose decreases R_0 by raising the active dopant in the highly doped source drain region. Poor HF clean tends to act as a dopant sink, preventing them from migrating toward the channel. Thus it tends to make underlapped transistors and raises σ parameter.

In a second step, within-wafer variability has been investigated on 14 nm FD-SOI technology. Monte Carlo, forward and backward propagation of variance have been conducted in order to model this variability. It has been shown that linear drain current variability is slightly underestimated. BPV and direct extraction showed close results in term of linear drain current variability however corresponding model parameter variability yield different results. It has thus been suggested that local variability and channel length variability are responsible for these discrepancies (that are not properly taken into account using direct extraction or BPV). This interpretation has been reinforced by the fact that Monte Carlo draws used to forward propagate the model parameter variability extracted using BPV and direct extraction gives the same results than FPV. This leads to infer that the discrepancy does not come from a violation of normality and linear local approximation hypothesis. In order to verify that channel length and local variability are responsible for observed discrepancies between measurements and model, their impact on the model has been assessed using synthetic data and showing that it induces errors and can thus explain it.

Chapter 5 introduces the procedure to build and calibrate polynomial formulas that relate process and model parameters (the first stage of our PCM depicted in Figure 6-2). Since process parameters are numerous and some of them are irrelevant depending on the considered model parameters, the task of building polynomial models faces two issues: i) the problem is ill-posed and ii) relevant variables should be selected. These issues are addressed using appropriate statistical method like stepwise regression, LASSO and LARS. It has been shown using synthetic data that these methods are able to perform variable selection with ill-posed problems and noisy observations.

The procedure has been applied on TCAD simulated DOE in order to test its reliability. Accurate results have been obtained. TCAD simulated within-wafer variability has been modeled using the PCM, showing a good agreement. Process parameters have then been ranked regarding their contribution to drain current variability. The model showed that Tepi and Wsp are mainly responsible for I_{Dlin} and I_{Dsat} variability. Thus we suggested optimizing these parameters in order to draw the best benefits in term of drain current variability.

In order to ensure the robustness of model building process, resources must meet specific requirements in term of amount and uncertainty of measurements. The limitations of the approach considering these requirements have been discussed. We have shown that the model construction is compromised by the impact of noise and local variability if only a single transistor is measured with short measurement time. However, we have shown that this problem can easily be overcome using array of transistors. We recommend using 20x20 transistors array, without increasing the measurement time in order to reach a proper noise level. Of course increasing the measurement time would increase the PCM construction robustness as well.

6.2 Application and perspectives

To sum up, this work is a feasibility study about PCM construction and shows how to benefit from extensive model parameters extractions in order to speed up the development process. We have shown that, with very little investments, the approach yields valuable results. Indeed, only few measurement points have been used instead of full I_D-V_G traditionally used for model calibration and no specific test structure. Every algorithm has been run using a laptop with average processing power associated with the flexible but rather slow Matlab software. Based on that, inferences about process effect on electrical performance have been drawn and PCM have been build based on TCAD simulated DOE. The PCM has been able to provide guidelines in order to optimize drain current variability.

The quality and quantity of drawn benefits are proportional to the amount of invested resources. In fact, there is a smooth trade-off between robustness and flexibility of the model and the required amount of resources to invest. In the following sections we investigate the potential benefits that could be drawn using PCM with some advanced features.

6.2.1 Optimizing the process flow

Variability optimization using PCM has been investigated on silicon, TCAD and synthetic data. However it has been suggested that this procedure can be applied to optimize performance and variability at the same time, provided that the model is well calibrated. In order to achieve this, we propose here, as an application, a general procedure aimed at optimizing performances and variability. This procedure also enables calibrating TCAD simulation tool. Description of the procedure is shown in Figure 6-3.

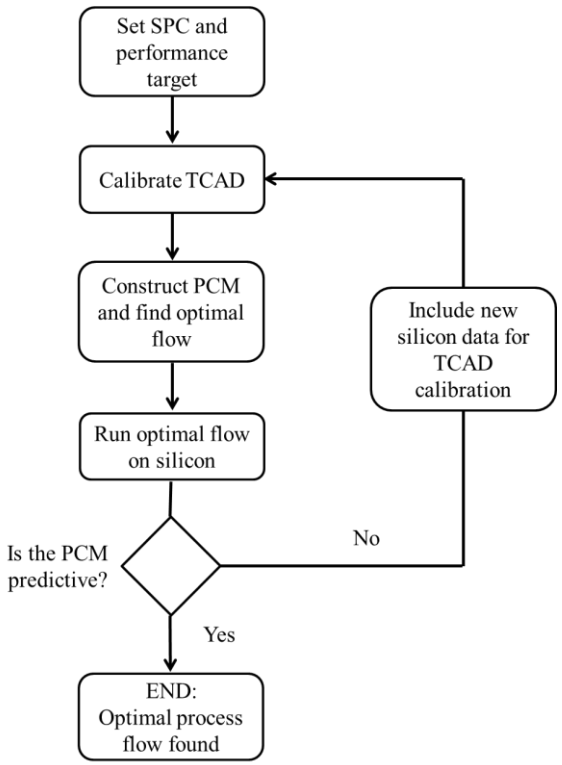


Figure 6-3: Flow chart of performance and variability optimization procedure

Figure 6-3 shows the flow chart of performance and variability optimization procedure. First, it consists in setting the target for the device in terms of electrical performance and variability. Then

TCAD simulations must be calibrated. This procedure can be done using the compact model and its extraction procedure. It is detailed in Figure 6-4.

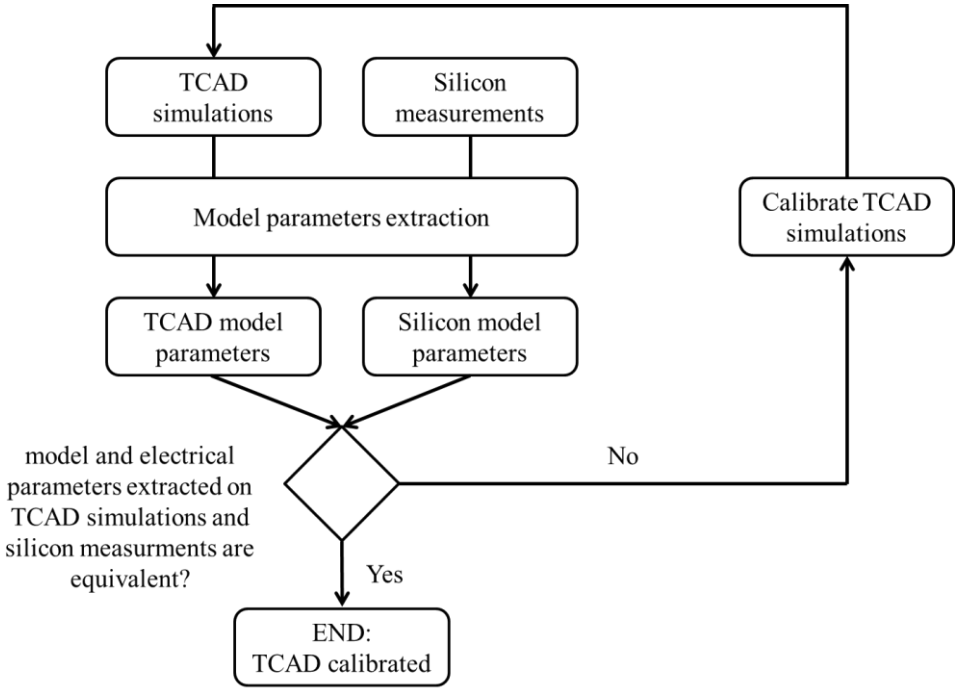


Figure 6-4: Flow chart of TCAD calibration procedure using compact model

In this procedure, we extract model parameters using TCAD and silicon measurements. Comparing electrical and model parameters of TCAD and silicon devices enable assessing the accuracy of TCAD calibration. If TCAD is not properly calibrated, then the mismatch between TCAD and silicon model parameters can indicate how to re-calibrate TCAD properly. For example if there is a good match between every TCAD and silicon model parameters with the exception of $\mu_0 C_{ox}$, then TCAD mobility model and/or equivalent oxide thickness should be investigated for TCAD calibration.

As soon as TCAD is calibrated, the PCM should be constructed (following Figure 6-3 flow chart). PCM construction procedure is detailed in Figure 6-5 flow chart. The procedure consists in simulating a DOE, extracting model parameters from simulated drain currents and building the PCM following instructions detailed in Chapter 5. Optimal process flow is then found using the PCM and an optimization algorithm. Relevance of the results must be checked just afterward. Indeed, if we considered the PCM build in Chapter 5, we can see that R_0 is linearly proportional to the implant dose. Thus if we would like to optimize the process flow such that it maximizes the drain current, then the solution would suggest to increase implanted dose indefinitely such that R_0 is minimized. In practice we know that there is a maximum dopant concentration above which no gain in access resistance can be expected, since there is a saturation effect in dopant concentration. In other words, the PCM domain of validity is too narrow and optimal solutions found are not bounded.

In order to correct this flaw, the device must be investigated when process parameters reach extreme values. For this example, simulations must be run with high enough implant doses in order to capture the concentration saturation effect. When the new DOE is designed, the simulations and PCM construction procedure must be run before optimizing the process flow once again. This loop should be repeated until a physically relevant and bounded solution is found for the process flow.

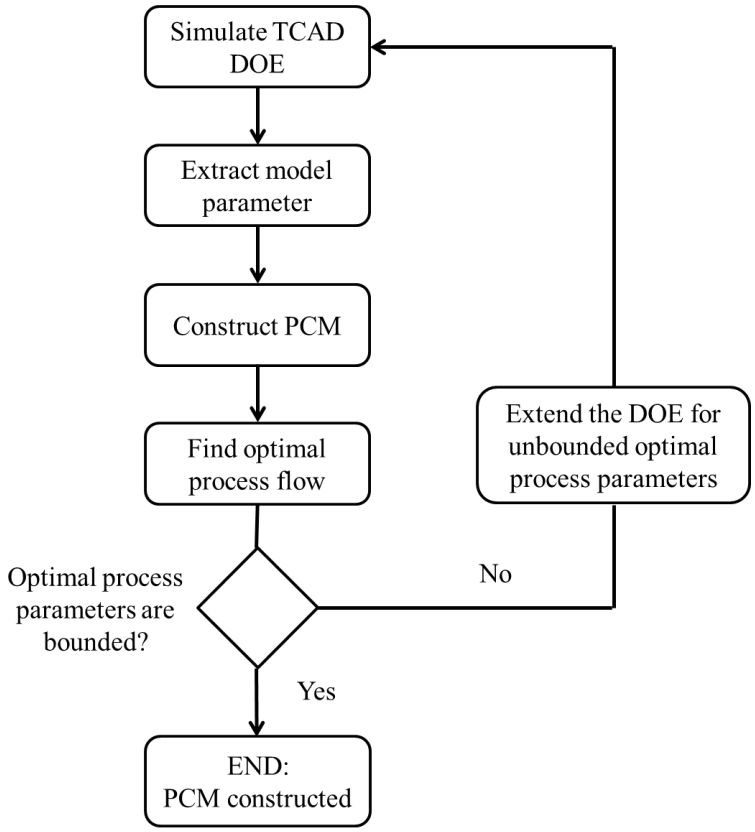


Figure 6-5: Flow chart of PCM construction and calibration

Following Figure 6-3 flow chart, a last check must be performed in order to determine if the process flow found by optimization is actually the optimal one. This test consists in running the optimal process flow on silicon. Results found using PCM and silicon measurements should be compared. If the model does not match silicon, it means that either TCAD is badly calibrated (considering this new process flow) or PCM is not predictive enough. The whole procedure should be run once again focusing now on this new process flow for model and TCAD calibration.

The entire procedure as shown in Figure 6-3 is iterative and little iterations could be required in order to come up with a consistent model and optimal process flow. It only requires to process one wafer per iteration, making the approach very cost effective. Moreover, a fully automated procedure can make the iteration very quick (in the order of magnitude of few hours). The only step that cannot be automated is TCAD calibration since it requires expert knowledge. However insight provided by model parameter extraction can greatly ease this task.

6.2.2 Advanced feature for future PCM studies

As a perspective, we propose here some guidelines in order to improve the approach developed in this work and draw full benefits of the technique. Figure 6-6 shows the PCM scheme like the one presented in introduction (see Figure 1-3). However here we have added advanced features that could be investigated in future studies.

Advanced features include a new type of model to link process and model parameters. It can be either Feed Forward Neural Network (FFNN) **[156][186][187]**, Support Vector Machine (SVM) **[188]** or simply user defined nonlinear function. In this study we used linear polynomial formula, but we have seen that it has the drawback to not include second order and cross effects.

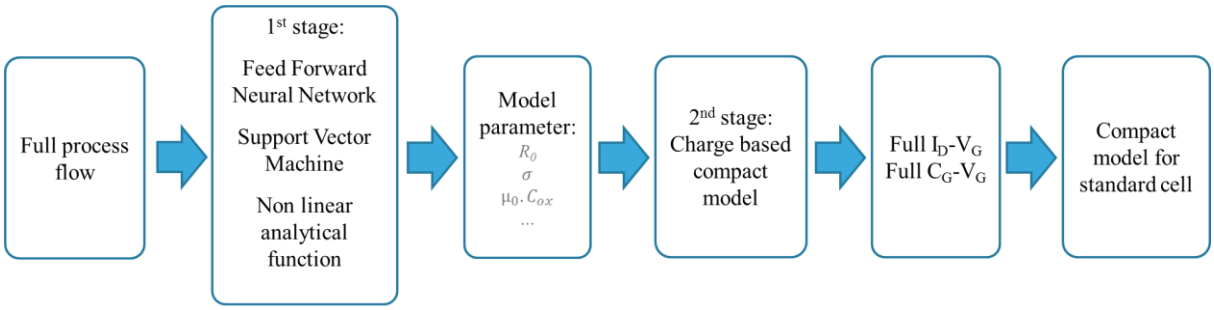


Figure 6-6: Flow chart of a three-stage PCM for standart cell

Moreover, nonlinear process dependence like sigma vs Tspike (see Figure 5-20), where tricky to model with linear polynomial. Finally, in order to find the optimal process flow, we have seen that the PCM should be valid for large process parameters variations and thus account for nonlinear relations between model and process parameters (such as saturation of the doping concentration for very high implant doses). This cannot be accounted for using linear polynomial but nonlinear formulas can account for it.

FFNN and SVM can also account for it. In addition these methods can deal with discontinuous parameters (e.g. Boolean variables). As a consequence it can deal with some shifts in the process flow (suppression or addition of steps, tool change). Considering FFNN, it should be noted that, even though this method is less transparent compared with polynomials, it is actually very easy to handle and to train. As well, it is very powerful in modeling complex systems, no matter its nature. This is why it is often called "universal approximation method". Today, artificial neural network find an increasingly large number of application, going from facial or speech recognition to game-playing and decision making, to medical diagnosis, just to name a few.

Aforementioned advanced features also includes a more flexible and accurate compact model. I suggested in Figure 6-6 using a charge based compact model but it can be surface potential based. The main guideline I would provide is to use a compact model with parameters that have a clear and elementary physical meaning. This asset would greatly simplify the PCM first stage and makes the whole PCM much more robust. In addition we can hope modeling the full I_D-V_G characteristic, even the C_G-V_G characteristic. Of course the main limitation is to have a limited number of model parameters (around 10) in order to be able to extract them with few measurements.

Finally, the last feature to be investigated is to extend the model by adding a third stage. This stage would take the transistor and standard cell electrical characteristic as input and output respectively (e.g. SNM for SRAM or switching speed for ring oscillators). This third step can be extremely valuable. Indeed in this thesis we have focused on optimizing I_{Dlin} and I_{Dsat} performances and variability. Targets to be reached (in term of drain current) are set so that it ensures the circuit functionalities. However, with a PCM able to model standard cell performances, it would be possible to directly optimize standard cell performances. This approach would provide more freedom and a broader range of solutions in term of optimal process flow. To another extend, it would be also possible to consider the effect of layout with this kind PCM. Thus optimization would not be limited to find the optimal process flow but also the optimal circuit layout. This kind of global optimization procedure would yield highly value-added solutions.

Figure 6-7 exposes a flow for PCM construction with advanced features that are worth investigating in future studies.

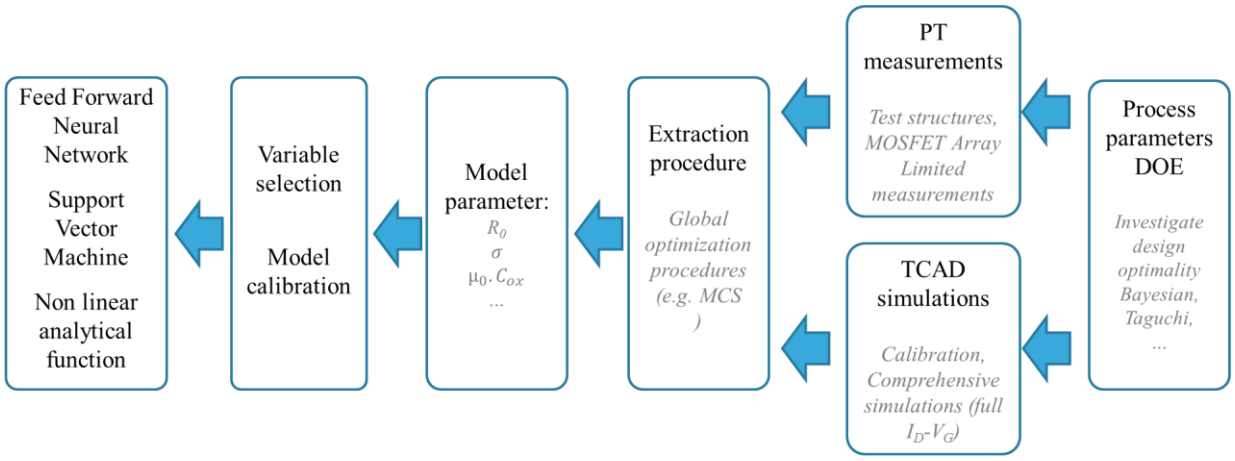


Figure 6-7: Simplified PCM construction flow with advanced feature

In this figure, the construction flow starts by building a proper DOE aimed at investigating process parameters effects. Depending on the model we are trying to buit (especially for the PCM first stage), different type of DOE can be used. Techniques to build optimal designs have been widely investigated. Common criterions (based on Fisher Information Matrix) to build optimal designs have been introduced by Wald (1943), Elving (1952), Kiefer (1959) and Kiefer (1975). A large and comprehensive literature can be found on this subject **[199]**. Choosing the right DOE minimizes the chance to build inaccurate PCMs and increases its robustness. For example, in order to build a second order polynomial model, a central composite design of experiment would be preferred to the DOE exposed in Table 5-4, that was used for linear polynomial model.

Another feature that would improve the technique is to use global optimization methods in order to improve the extraction procedure. In our approach we used a trust-reflective-region with conjugate gradient algorithm. It is very efficient considering our problem but it cannot ensure finding the global optimum of the problem, especially if the first guess is bad. In this work, we have circumvented this issue by using first guess found with linear least square fit. This approach ensures to have a first guess close enough to the global optimum. Moreover the extraction robustness has been tested in a great extent. So considering our case, the optimization method works fine. But if a more accurate compact model is used for which no first guess can be provided before running the optimization step, then a global optimization algorithm could be beneficial. A wide range of solutions exist and some of them have been investigated in literature for parameter extraction **[133][134][200][201]**. Different algorithms have been tested during this thesis (like genetic algorithm and levenberg-marquardt), and we recommend using derivative based optimization algorithm since compact models are continuous, derivable and fast-to-compute. These methods appear to be quicker and more accurate. In this perspective we can mention the Multi-levels Coordinate Search (MCS) as a global optimization alternative to our approach **[202]**. More information about global optimization can be found in literature **[203]**.

If one was to use FFNN, SVM or nonlinear analytical formula for the PCM first stage, then specific methods should be used for variable selection and model calibration. FFNN and SVM do not explicitly requires variable selection, but it can improve its efficiency for training and computing. Literature reports multiple methods to perform variable selection, in accordance with these methods **[204]-[208]**. In case one wants to test nonlinear model, other methods to perform variable selection are more appropriate **[180]-[186]**. These methods also serve at calibrating nonlinear functions.

6.2.3 Unexplored application

In term of unexplored applications we have already mentioned the possibility to optimize standard cell instead of transistors performances. Some other application can be suggested. Up to now, we have described a technique to finding the optimal process flow. However, during the process, performance and variability can be impacted by calibration drift of processing tools. In other word, tools calibration can drift with time and slightly modify process parameters mean values over the wafer. In the end, there will be some discrepancies between the optimal process flow and the one that was actually processed. The same problem can be observed at die scale. Indeed, process parameters are not uniformly distributed over the wafer, due to within wafer variability. Often, there is a wafer signature (e.g. radial or linear gradient) of process parameters dispersion. Thus, the process observed at die scale can be different from the optimal process flow. In order to counteract this issue, a study has implemented in situ process adjustment, in order to reduce the effect of process drift **[157]**. The method consists in making real time, in line, process monitoring in order to estimate the process drift at die scale with respect to the optimal process flow. Then at some critical step of the flow, the process can be adjusted thanks to tool recalibration or wafer reorientation in order to counteract the effect of tool calibration drift and wafer signature of process parameters. This process requalification can be done using the PCM. At the critical process step, instead of optimizing the entire process flow as we suggested in previous application, only the remaining process steps would to be optimized, knowing the process history. This real time, in situ process optimization thanks to PCM would allow yield and performance maximization.

References

[1] U. Pirzada, “Intel Hints At New 2.5 Year ‘Tick Tock Tock’ Cadence – Confirms 14nm Kaby Lake Intermediary Platform and 10nm Cannonlake Delay to 2017” article available at : http://wccftech.com/intel-new-2-5-year-cadence-tick-tock-tock/, 2015.

[2] Victor Moroz, “Transition from Planar MOSFETs to FinFETs and its Impact on Design and Variability”, Berkley seminar, 2011.

[3] R. Sitte, S. Dimitrijev, H. B. Harrison, “Device parameter changes caused by manufacturing fluctuations of deep submicron MOSFET’s” IEEE Trans. Electron Devices, vol. 41, no. 11, pp. 2210-2215, 1994.

[4] R. Sitte, S. Dimitrijev, H. B. Harrison, “Sensitivity of 0.1 µm MOSFET’s to manufacturing fluctuations”, Electronic Letter, vol. 29, no. 15, pp. 1345-1346, 1993.

[5] F.-L. Yang, J.-R. Hwang, Y. Li, “Electrical characteristic fluctuation in Sub-45nm CMOS Devices”, IEEE Custom Integrated Circuits Conference, pp 691-694, 2006.

[6] S. K. Saha, “Compact MOSFET modeling for process variability-Awar VLSI circuit design”, IEEE Access, Vol. 2, pp. 104-115, 2014.

[7] C-H Hsiao, D-M Kwa, “Measurement and Characterization of 6T SRAM Cell Current”, IEEE Proc. of International Workshop on Memory Technology, Design, and Testing, 2005.

[8] J. Hu, J.-E. Park, G. Freeman, R. Wachnik, H.-S. P. Wong “Effective Drive Current in CMOS Inverters for Sub-45nm Technologies”, NSTI-Nanotech, pp. 829-832, 2008.

[9] M.-H. Na, E.J. Nowak, W. Haensch, J. Cai, “Effective Drive Current in CMOS Inverters”, IEEE Proc. on International Electron Device Meeting, pp. 121-124, 2002.

[10] J. P. Colinge, “Recent Advances and Trends in SOI CMOS Technology”, IEEE Proc. on Solid State Device Research Conference, pp. 935-942, 1996

[11] B. Yu, L. Chang, S. Ahmed, W. Haihong; S. Bell, Y. Chih-Yuh, C. Tabery, Chau Ho, Qi Xiang, Tsu-Jae King, J. Bokor, Chenming Hu, L. Ming-Ren; D. Kyser, “FinFET scaling to 10 nm gate length”, IEEE Proc. on IEDM, pp. 251 – 254, 2002.

[12] H.-S.P. Wong, K.K. Chan, Y. Taur, “Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel”, IEEE Proc. on IEDM, pp. 427 – 430, 1997

[13] J.P. Colinge, M. H. Gao, A. Romano-Rodriguez, H. Maes, C. Claeys, “Silicon-on-insulator 'gate-all-around device"', IEEE Proc. on IEDM, pp. 595 - 598, 1990

[14] L. Laurent, “TCAD Modeling of variability in 28-nm CMOS technologies based on Impedance Field Method”, M.S. thesis, Ecole Centrale de Lille, France, 2012

[15] K. Qian, “Variability Modeling and Statistical Parameter Extraction for CMOS Devices”, Ph.D. dissertation, Electrical Engineering and Computer Sciences University of California at Berkeley, 2015.

[16] C.-H. Lin, M. V. Dunga, D. D. Lu, A. M. Niknejad, and C. Hu, “Performance-Aware Corner Model for Design for Manufacturing,” IEEE Trans. Electron Devices, vol. 56, no. 4, pp. 595-600, 2009.

[17] M. Kanno, A. Shibuya, M. Matsumura, K. Tamura, H. Tsuno, S. Mori, Y. Fukuzaki, T. Gocho, H. Ansai, and N. Nagashima, “Empirical Characteristics and Extraction of Overall Variations for 65-nm MOSFETs and Beyond,” vol. 2, pp. 88–89, 2007.

[18] A. J. Strojwas, “Cost-Effective Variability Reduction Approaches to Enable Future Technology Nodes”, IEEE Proc. of SISPAD, 2010.

[19] M. Yakupov, D. Tomaszewski, “Analysis of Selected Methods for CMOS Integrated Circuit Design for Yield Optimization”, International Conference "Mixed Design of Integrated Circuits and Systems", pp.

71-76, 2011.

[20] X. Zhang, X. Bai "Process Variability-Induced Timing Failures — A Challenge in Nanometer CMOS Low-Power Design", IEEE Conference publication, pp. 159-162, 2008.

[21] B. Cheng, S. Roy, A. R. Brown, C. Millar, A. Asenov, "Evaluation of Intrinsic parameter fluctuation on 45, 32 and 22 nm technology node LP n-MOSFETs", IEEE Proc. on ESSDERC, pp. 47-50, 2008.

[22] a. Asenov, a. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," IEEE Trans. Electron Devices, vol. 50, no. 9, pp. 1837-1852, 2003.

[23] B. Cheng, S. Roy, G. Roy, F. Adamulema, and a Asenov, "Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells," Solid. State. Electron., vol. 49, no. 5, pp. 740-746, May 2005.

[24] A. Asenov, S. Kaya, and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," IEEE Trans. Electron Devices, vol. 49, no. 1, pp. 112-119, 2002.

[25] A. Asenov, B. Cheng, D. Dideban, U. Kovac, N. Moezi, C. Millar, G. Roy, A. Brown, and S. Roy, "Modeling and simulation of transistor and circuit variability 140 and reliability," in Custom Integrated Circuits Conference (CICC), 2010 IEEE, 2010, pp. 1-8.

[26] B. Bindu, B. Cheng, G. Roy, X. Wang, S. Roy, and a. Asenov, "Parameter set and data sampling strategy for accurate yet efficient statistical MOSFET compact model extraction," Solid. State. Electron., vol. 54, no. 3, pp. 307-315, Mar. 2010.

[27] L. Rahhal, A. Bajolet, C. Diouf, A. Cros, J. Rosa, N. Planes, G. Ghibaudo, "New methodology for drain current local variability characterization using Y function method", IEEE Proc. of ICMTS, pp. 99-103, 2013.

[28] L. Rahhal, A. Bajolet, J.-P. Manceau, J. Rosa, S. Ricq, S. Lassere, G. Ghibaudo "Mismatch trends in 20nm gate-last bulk CMOS technology", IEEE Proc. of ULIS, pp. 133 – 136, 2014.

[29] L. Rahhal, A. Bajolet, J.-P. Manceau, J. Rosa, S. Ricq, S. Lassere, G. Ghibaudo, "A comparative mismatch study of the 20 nm Gate-Last and 28 nm Gate-First bulk CMOS technologies", Solid State Electronic. Vol. 108, pp. 53-60, 2015.

[30] R. H. Kingston and S. F. Neustadter, "Calculation of the Space Charge, Electric Field, and Free Carrier Concentration at the Surface of a Semiconductor", J. Appl. Phys. 26, 718 (1955).

[31] C. C. Hu, "MOS Transistor", in Modern Semiconductor Devices for Integrated Circuits, 1st Ed., Prentice Hall, 2010, ch. 5, sec. 5.3, pp. 162.

[32] S. M. Sze, Physics of Semiconductor Devices, 2nd Ed. New York, USA, John Wiley & Sons, 1981.

[33] R. S. Muller, T. I. Kamins, "Metal semiconductor contact", in Device electronics for integrated circuits, 2nd Ed. New York, USA, John Wiley & Sons, 1986, ch. 3, sec. 2, pp.133-134.

[34] D. Garetto, D. Rideau, C. Tavernier, Y. Leblebiciand, A. Schmid, H. Jaouen, "Advanced physics for simulation of ultrascaled devices with UTOXPP Solver", Nanotechnology, Vol. 2, Ch. 9, pp.854, 2011.

[35] H. K. Lim, J; G. Fossum, "Threshold Voltage of Thin-Film Silicon-on-Insulator (SOI) MOSFET's, IEEE Trans. Elec. Dev., Vol. ED30, no. 10, pp. 1244-1251, 1983.

[36] Lundström cours EE-612, Lecture 25: "SOI Electrostatics". Available at : https://nanohub.org/resources/6016/download/2008.12.04-ece612-l25.pdf.

[37] F. Andrieux, « Transistor CMOS decananométriques à canaux contraints sur silicium massif ou sur SOI – fabrication, caractérisation et étude du transport », Ph.D dissertation, CEA-LETI, Institut national polytechnique de Grenoble, France, 2005.

[38] V. P. Trivedi, J. G. Fossum, “Scaling Fully Depleted SOI CMOS”, IEEE Trans. Elec. Dev., Vol. 50, no. 10, pp. 2095-2103, 2003.

[39] V. P. Trivedi, J. G. Fossum, W. Zhang, “Threshold voltage and bulk inversion effects in nonclassical CMOS devices with doped ultra-thin bodies”, Solide states electronics, Vol. 51, pp. 170-178, 2007.

[40] C. T. Lee, K. K. Young, “Submicrometer Near-Intrinsic Thin-Film SOI Complementary MOSFET’s”, IEEE Trans. Elec. Dev., Vol. 36, no. 11, pp. 2537-2547, 1989.

[41] Q. Chen, E. M. Harrell, and J. D. Meindl, “A physical short-channel threshold voltage model for undoped symmetric double-gate MOSFETs,” IEEE Trans. Elec. Dev., vol. 50, no. 7, pp. 1631–1637, Jul. 2003.

[42] J. Lacord, J. L. Huguenin, T. Skotnicki, G. Ghibaudo, F. Boeuf, “and Efficient MASTAR Threshold Voltage and Subthreshold Slope Models for Low-Doped Double-Gate MOSFET”, IEEE Trans. Elec. Dev., Vol. 59, no. 9, 2012.

[43] S. Eminente, S. Cristoloveanu, R. clerc, A. Ohata, G. Ghibaudo, “Ultra-thin fully depleted SOI MOSFETs: Special charge properties and coupling effects”, Solid-State Electronics, Vol. 51, pp. 239-244, 2007.

[44] S. Burignat, D. Flandre, M. K. Md Arshad, V. Kilchytska, F. Andrieu, O. Faynot, J.-P. Raskin, “Substrate impact on threshold voltage and subthreshold slope of sub-32 nm ultra-thin SOI MOSFETs with thin buried oxide and undoped channel”, Solid State Electronics, Vol. 54, pp. 213-219, 2010.

[45] S-I. Takagi, A. Toriumi, M. Iwase, H. Tango, “On the Universality of Inversion Layer Mobility in Si MOSFET's: Part I-Effects of Substrate Impurity Concentration”, IEEE Trans. Elec. Dev., Vol. 41, no. 12, 1994.

[46] F. Boeuf, G. Ghibaudo, and T. Skotnicki, “Impact of Coulomb scattering on the characteristics of nanoscale devices", International Conference on Solid State Devices and Materials (SSDM), Sendai, Japan, 2009.

[47] C. Fenouillet-Beranger, S. Denorme, P. Perreau, C. Buj, O. Faynot, F. Andrieu, L. Tosti, S. Barnola,

T. Salvetat, X. Garros, M. Cassé, F. Allain, N. Loubet, L. Pham-NGuyen, E. Deloffre, M. Gros-Jean,

R. Beneyton, C. Laviron, M. Marin, C. Leyris, S. Haendler, F. Leverd, P. Gouraud, P. Scheiblin, L. Clement, R.Pantel, S. Deleonibus, T. Skotnicki,“FD-SOI devices with Thin BOX and Ground plane integration for 32nm node and below”, Proc. in European Solid State Device Research Conference (ESSDERC), pp206-209, 2008

[48] G. Hiblot, « Compact modeling of MOSFET transistors with III-V channels and thin films for advanced CMOS applications», Ph.D dissertation, IMEP-LAHC, Université de Grenoble, France, 2015.

[49] J. Koga, S. Takagi, and A. Toriumi, “A comprehensive study of MOSFET electron mobility in both weak and strong inversion regimes", Electron Devices Meeting, 1994. IEDM '94. Technical Digest., International, Dec 1994, pp. 475-478.

[50] M. J. Sherony, L. T. Su, J. E. Chung, D. A. Antoniadis, “SOI MOSFET effective channel mobility”, IEEE Trans. Elec. Dev., Vol. 41, no. 2, pp. 276-278, 1994

[51] S. Takagi, A. Toriumi, M. Iwase, H. Tango, “On the universality of inversion layer mobility in Si MOSFET’s: Part I-Effects of substrate impurity concentration”, IEEE Trans. Elec. Dev., Vol. 41, no. 12, pp. 2357, 1994.

[52] D. Rideau, Y. M. Niquet, O. Nier, A. Cros, J.P. Manceau, P. Palestri, D. Esseni, V. H. Nguyen, F. Triozon, J.C. Barbé, I. Duchemin, D. Garetto, L. Smith, L. Silvestri, F. Nallet, R. Clerc, O . Weber, F. Andrieu, E. Josse, C. Tavernier, H. Jaouen, "Mobility in High-K Metal Gate UTBB-FD-SOI Devices: from NEGF to TCAD perspectives", Proc. on International Electron Device Meeting (IEDM), pp. 12.5.1-12.5.4, 2013.

[53] G. Ghibaudo, M. Mouis, L. Pham-Nguyen, K. Bennamane, I. Pappas, A. Cros, G. Bidal, D. Fleury, A. Claverie, G. Benasayag, P-F. Fazzini, C. Fenouillet-Beranger, S. Monfray, F. Boeuf, S. Cristoloveanu, T. Skotnichi, N. Collaert, "Electrical transport characterization of nano CMOS devices with ultra-thin silicon film", Ext. Abs. 9th international workshop on junction technology, pp. 58-63, 2009.

[54] V. Barral, T. Poirroux, D. Munteanu, J-L. Autran, S. Deleonibus, "Experimental investigation on the quasi-ballistic transport: Part II-backscattering coefficient extraction and link with the mobility.", IEEE Trans. Elec. Dev., Vol. 56, no.3, pp. 420-430, 2009.

[55] I. Pappas, G. Ghibaudo, C. A. Dimitriadis, C. Fenouillet-Beranger, "Backscattering coefficient and drift-diffusion mobility extraction in short channel MOS devices", Solid States Electronics, Vol. 53, pp. 54-56, 2009.

[56] S. Guarnay, F. Triozon, S. Martinie, Y. M. Niquet, A. Bournel, "Monte Carlo study of effective mobility in short channel FD-SOI MOSFETs", Proc. of SISPAD international conference, 2014

[57] D. Fleury, G. Bidal, A. Cros, F. Boeuf, T. Skotnicki, G. Ghibaudo, "New Experimental insight into ballistic of transport in strained bulk MOSFETs", Proc. of Symposium on VLSI technology, pp. 16-17, 2009.

[58] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, S. Mukhopadhyay, "Magnetoresistance mobility characterization in advanced FD6SOI n-MOSFETs", Solid States Electronics, Vol. 10, pp. 229-235, 2015.

[59] M. Zilli, P. Palestri, D. Esseni, L. Selmi, "On the experimental determination of channel backscattering in nano MOSFETs.", IEDM Tech Digest, pp.105, 2007.

[60] K. Huet, J. Saint-Martin, A. Bournel, S. Galdin-Retailleau, P. Dollfus, G. Ghibaudo, M. Mouis, "Monte Carlo study of apparent mobility reduction in nano-MOSFETs", Proc of ESSDERC, pp. 382-385, 2007

[61] E. J. Ryder, "Mobility of Holes and Electrons in High Electric Fields", Physical Review, Vol. 90, no. 5, pp. 766-769, 1953.

[62] C. B. Norris, J. F.Gibbons, "Papers on Carrier Drift Velocities in Silicon at High Electric Field Strengths", IEEE Trans. Elec. Dev., Vol. 14, no. 1, pp.37, 1967.

[63] C. Y. Duh, J. L. Moll, "Electron Drift Velocity in Avalanching Silicon Diodes", IEEE Trans. Elec. Dev., Vol. 14, no. 1, pp.46, 1967.

[64] V. Rodriguez, H. Ruegg, M-A. Nicolet, "Measurement of the drift velocity of holes in silicon at high-field strengths", IEEE Trans. Elec. Dev., Vol. 14, no. 1, pp.44, 1967.

[65] J. G. Ruch, "Electron Dynamics in Short Channel Field-Effect Transistors" , IEEE Trans. on Elec. Dev., Vol. 19, no. 5, pp. 652-654, 1972.

[66] J. Kim, J. Lee, Y. Yun, B-G. Park, J. D. Lee, H. Shin, "Extraction of Effective Carrier Velocity and Observation of Velocity Overshoot in Sub-40 nm MOSFETs", Journal of semiconductor technology and science, Vol. 8, no.2, pp.115-120, 2008.

[67] M. Lundstrom, “Elementary Scattering Theory of the Si MOSFET”, IEEE Elec. Dev. Lett., Vol. 18, no. 7, pp. 361-363, 1997

[68] Peizhen Yang, W.S. Lau, Seow Wei Lai, V.L. Lo, S.Y. Siah and L. Chan (2010). The Evolution of Theory on Drain Current Saturation Mechanism of MOSFETs from the Early Days to the Present Day, Solid State Circuits Technologies, Jacobus W. Swart (Ed.), ISBN: 978-953-307-045-2, InTech, DOI: 10.5772/6873. Available from: http://www.intechopen.com/books/solid-state-circuits-technologies/the-evolution-of-theory-on-drain-current-saturation-mechanism-of-mosfets-from-the-early-days-to-the-

[69] K. Natori, “Ballistic metal-oxide semiconductor field effect transistor”, Journal of Applied Physics, Vol. 76, no. 8, pp. 4879-4890, 1994.

[70] C. C. Hu, “MOS Transistor”, in Modern Semiconductor Devices for Integrated Circuits, 1st Ed., Prentice Hall, 2010, ch. 6, sec. 6.3.1, pp. 202.

[71] G. Ghibaudo, “Analytical modeling of the MOS transistor”, Phys. Stat. Sol., Vol. 113, pp. 223-239, 1989.

[72] G. Ghibaudo, “A simple model of the drain saturation voltage dependence with gate voltage for short channel MOSFETs”, Phys. Stat. Sol., Vol. 99, pp. K149-K153, 1987.

[73] L. Pham-Nguyen, C. Fenouillet-Beranger, A. Vandooren, A. Wild, G. Ghibaudo, S. Cristoloveanu, “Direct comparison of Si/High-K and Si/SiO2 channels in advances FD SOI MOSFETs”, Proc. of IEEE International SOI conference, pp. 25-26, 2008.

[74] M. Cassé, F. Rochette, N. Bhouri, F. Andrieu, D. K. Maude, M. Mouis, G. Reimbold, F. Boulanger, “Mobility of strained and unstrained short channel FD-SOI MOSFETs: new insight by magnetoresistance”, Proc. of Symposium on VLSI technology digest of technical papers, pp. 170-171, 2008.

[75] W. Chaisantikulwat, M. Mouis, G. Ghibaudo, C. Gallon, C. Fenouillet-Beranger, D.K. Maude, T. Skotnicki, S. Cristoloveanu, ”Magnetoresistance technique for mobility extraction in short channel FD-SOI transistors”, Proc. of IEEE ESSDERC, pp. 569-572, 2005.

[76] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G. Kim, G. Ghibaudo, “In depth characterization of electron transport in 12nm FD-SOI CMOS”, Solid States Electronics (2015), http://dx.doi.org/10.1016/j.sse.2015.02.012.

[77] S. Morvan, F. Andrieu, M. Cassé, O. Weber, N. Xu, P. Perreau, J. M. Hartmann, J. C. Barbé, J. Mazurier, P. Nguyen, C. Fenouillet-Beranger, C. Tabone, L. Tosti, L. Brévard, A. Toffoli, F. Allain, D. Lafond, B. Y. Nguyen, G. Ghibaudo, F. Boeuf, O. Faynot, T. Poirroux, “Efficiency of mechanical stressors in planar FD-SOI n and p MOSFETs down to 14nm gate length”, Proc. of Symposium on VLSI technology digest of technical papers, pp. 111-112, 2012.

[78] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G.-T. Kim, “Low temperature characterization of mobility in advanced FD-SOI n-MOSFETs under interface coupling conditions”, Proc on ULtimate Iintegrated on Silicon conference, pp. 61-64, 2014.

[79] S. R. Hofstein, F. P. Heiman, “Insulated-gate field effect transistor”, Proceedings of the IEEE, Vol. 51, no. 9, pp. 1190-1202, 1963.

[80] G. Merckel, J. Borel, N. Z. Cupcea, “An Accurate Large-Signal MOS Transistor Model for se in Computer-Aided Design”, IEEE Trans. Elec. Dev., Vol. ED19, no. 5, pp. 681-690, 1972.

[81] P. I. Suciu, R. L. Johnston, “Experimental Derivation of the Source and Drain Resistance of MOS

Transistors", IEEE Trans. Elec. Dev. Vol. ED27, no. 9, pp. 1846-1848, 1980.

[82] B. Cabon-Till, G. Ghibaudo, S. Cristoloveanu, "Influence of source drain series resistance on MOSFET Field-effect mobility", Electronics Letters, Vol. 21, no. 11, pp. 457-458, 1985.

[83] G. J. Hu, C. Chang, Y. T. Chia, "Gate-Voltage-Dependent Effective Channel Length and Series Resistance of LDD MOSFET's", IEEE Trans. Elec. Dev. Vol. ED34, no. 12, pp. 2469-2475, 1987.

[84] K. K. Ng, W. T. Lynch, "Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFET's", IEEE Trans. Elec. Dev. Vol. ED33, no. 7, pp. 965-972, 1986.

[85] V. G. K. Reddi, C. T. Sah, "Source to Drain Resistance Beyond Pinch-Off in Metal-Oxide-Semiconductor Transistors (MOST)", IEEE Trans. Elec. Dev. Vol. 12, no. 3, pp. 139-141, 1965.

[86] F. Monsieur, Y. Denis, D. Rideau, J. Lacord, V. Quenette, G. Gouget, C. Tavernier, H. Jaouen, 'The importance of the spacer region to explain short channels mobility collapse in 28 nm Bulk and FD-SOI technologies', IEEE Proc. on ESSDERC 2014.

[87] K. Y. Lim and X. Zhou, "A Physically-Based Semi-Empirical Series Resistance Model for Deep-Submicron MOSFET I–V Modeling", IEEE Trans. Elec. Dev., Vol. 47, no. 6, pp. 1300-13012, 2000.

[88] B. J. Sheu, C. Hu, P. K. KO and F. C. Hsu, "Source-and-Drain Series Resistance of LDD MOSFET's", IEEE Elec. Dev. Lett., Vol. EDL-5, no. 49, pp. 365-367, 1984.

[89] S. D. Kim, C. M. Park, J. C. S. Woo, "Advanced Model and Analysis of Series Resistance for CMOS Scaling Into Nanometer Regime—Part I: Theoretical Derivation", IEEE Trans. Elec. Dev. Vol. 49, no. 3, pp. 457-466, 2002.

[90] Y. Taur, "MOSFET channel length: extraction and interpretation", IEEE Trans. Elec. Dev., Vol. 47, no. 1, pp. 160-170, 2000

[91] J. Kim, J. Lee, I. Song, Y. Yun, J. D. Lee, B-G. Park, H. Shin, "Accurate extraction of effective channel length and source/drain resistance on ultrashort channel MOSFETs by iteration method", IEEE Trans. Elec. Dev., Vol. 55, no. 10, 2008.

[92] M.F. Hamer, "First-order parameter extraction on enhancement silicon MOS transistors", IEEE Proc., Vol. 133, Pt. 1, no. 2, pp. 49-54, 1986.

[93] Q. Chen, E. M. Harrell, and J. D. Meindl, "A physical short-channel threshold voltage model for undoped symmetric double-gate MOSFETs," IEEE Trans. Elec. Dev., vol. 50, no. 7, pp. 1631–1637, Jul. 2003.

[94] Y. P. Tsividis, C. McAndrew, "Operation and Modeling of the MOS Transistor", Mc Graw-Hill Book Company, New York, p.155, (1987).

[95] H.S.P. Wong, M.H. White, T.J. Krutsick, R.V. Booth "Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFETs" , Solide-State Electronics, Vol. 30, no. 10, pp 953-968, 1987.

[96] J. Lacord, J. L. Huguenin, T. Skotnicki, G. Ghibaudo, F. Boeuf, "and Efficient MASTAR Threshold Voltage and Subthreshold Slope Models for Low-Doped Double-Gate MOSFET", IEEE Trans. Elec. Dev., Vol. 59, no. 9, 2012.

[97] F. Balestra, I. Hafez, G. Ghibaudo, "A new method for the extraction of MOSFET parameters at ambient and liquid helium temperatures", Journal de physique, Colloque C4, supplement au no. 9, Tome 49, pp 817-820, 1988.

[98] A. Ortiz-Conde, F.J. Garcia Sanchez, J.J. Liou, A. Cerdeira, M. Estrada, Y. Yue, "A review of recent MOSFET threshold voltage extraction methods", Microelectronics reliability, Vol. 42, pp. 583–596, 2002.

[99] C. C. McAndrew and P. A. Layman, "MOSFET Effective Channel Length, Threshold Voltage, and Series Resistance Determination by Robust Optimization", IEEE Trans. Elec. Dev., Vol. 39, pp. 2298-2311, 1992.

[100] K. K. Ng, J. R. Brews, "Measuring the effective channel length of MOSFETs", IEEE Circ. And Dev. Mag., Vol. 6, no. 6, pp. 33-38, 1990.

[101] C. C. McAndrew, P. A. Layman, "MOSFET effective channel length, threshold voltage and series resistance determination by robust optimization", IEEE Trans. Elec. Dev., Vol. 39, no. 10, pp. 2298-2311, 1992.

[102] P. I. Suciu, R. L. Johnston, "Experimental Derivation of the Source and Drain Resistance of MOS Transistors", IEEE Trans. Elec. Dev. Vol. ED27, no. 9, pp. 1846-1848, 1980.

[103] B. Cabon-Till, G. Ghibaudo, S. Cristoloveanu, "Influence of source drain series resistance on MOSFET Field-effect mobility", Electronics Letters, Vol. 21, no. 11, pp. 457-458, 1985.

[104] C. Hao, B. Cabon-Till, S. Cristoloveanu and G. Ghibaudo, 'Experimental determination of short-channel MOSFETs parameters', Solid-State Electronics, Vol. 28, no. 10, pp. 1025-1030, 1985.

[105] F. Monsieur, Y. Denis, D. Rideau, J. Lacord, V. Quenette, G. Gouget, C. Tavernier, H. Jaouen, 'The importance of the spacer region to explain short channels mobility collapse in 28 nm Bulk and FD-SOI technologies', IEEE Proc. on ESSDERC 2014.

[106] G. J. Hu, C. Chang, Y. T. Chia, "Gate-Voltage-Dependent Effective Channel Length and Series Resistance of LDD MOSFET's", IEEE Trans. Elec. Dev. Vol. ED34, no. 12, pp. 2469-2475, 1987.

[107] K. Terada and H. Muta, "A new method to determine effective MOSFET channel length", Japan. J. Appl. Phys., Vol. 18, no. 5, pp. 953, 1979.

[108] J. G. J. Chem, P. Chang, R. F. Motta, and N. Godinho, "A new method to determine MOSFET channel length", IEEE Elec. Dev. Lett., Vol. EDL-l, no. 9, pp. 170, 1980.

[109] F. H. De La Moneda, H. N. Kotecha and M. Shatzkes, "Measurement of MOSFET Constants", IEEE Elec. Dev. Lett., Vol. EDL-3, no.1 , pp. 10-12, 1982.

[110] K. L. Peng and M. A. Afromowitz, "An Improved Method to Determine MOSFET Channel Length", IEEE Elec. Dev. Lett., Vol. EDL- 3, no. 12, pp. 360-362, 1982.

[111] B. J. Sheu, C. Hu, P. K. KO and F. C. Hsu, "Source-and-Drain Series Resistance of LDD MOSFET's", IEEE Elec. Dev. Lett., Vol. EDL-5, no. 49, pp. 365-367, 1984.

[112] K. –L. Peng, S. –Y. Oh. M. A. Afromowitz and J. L. Moll, "Basic Parameter Measurement and Channel Broadening Effect in the Submicrometer MOSFET", IEEE Elect. Dev. Lett., Vol. EDL-5, no. 11, pp. 473-475, 1984.

[113] J. Whitfield, "A Modification on 'An Improved Method to Determine MOSFET Channel Length", IEEE Elect. Dev. Lett., Vol. EDL-6, no. 3, pp. 109-110, 1985.

[114] L. Chang and J. Berg, "A Derivative Method to Determine a MOSFET's Effective Channel Length and Width Electrically", IEEE Elec. Dev. Lett., Vol. EDL-7, no. 4, pp. 229, 1986.

[115] F. Balestra, I. Hafez, G. Ghibaudo, "A new method for the extraction of MOSFET parameters at ambient and liquid helium temperatures", Journal de physique, Colloque C4, supplement au no. 9,

Tome 49, pp 817-820, 1988.

[116] F. Balestra, I. Hafez, G. Ghibaudo, “Modeling of electron mobility in silicon MOS inversion and accumulation layers at liquid helium temperature”, Electronics letters, Vol. 26, no. 19, pp 1633-1635, 1990.

[117] A. Cros, S. Harrison, R. Cerutti, P. Coronel, G. Ghibaudo, H. Brut, “New extraction method for gate bias dependent series ressitance in nanometric double gate transistors”, Proc. IEEE ICMTS, Vol. 18, 2005.

[118] D. Fleury, A. Cros, H. Brut, G. Ghibaudo, “New Y –Function-Based Methodology for Accurate Extraction of Electrical Parameters on Nano-Scaled MOSFETs”, Proc. IEEE ICMTS, 2008.

[119] D. Fleury, A. Cros, G. Bidal, J. Rosa, G. Ghibaudo, “A New Technique to Extract the Source/Drain Series Resistance of MOSFETs”, IEEE Elec. Dev. Lett., Vol. 30, no. 9, pp 975-977, 2009.

[120] N. Subramanian, G. Ghibaudo, M. Mouis, “Parameter Extraction of Nano-Scale MOSFETs Using Modified Y Function Method”, Proc. of IEEE ESSDERC, 2010.

[121] Y.Taur et al. “A New “Shift and Ratio” Method for MOSFET Channel-Length Extraction”, IEEE Elect. Dev. Lett., EDL-13(5), p. 267, 1992.

[122] S. Biesemans, M. Hendriks, S. Kubicek and K. De Meyer, “Accurate determination of Channel Length, Series Resistance and Junction Doping Profile for MOSFET 170ptimization in deep submicron technologies”, Proc. Symp. VLSI Technology Tech. Digest, p. 166, 1996.

[123] F. J. G. Sanchez, A. Ortiz-Conde, A. Cerdeira, M. Estrada, D. Flandre, J. J. Liou, “A method to extract mobility degradation and total series resistance of fully depleted SOI MOSFETs”, IEEE Trans. Elec. Dev., Vol 49, no. 1, 2002.

[124] K. O. Jeppson, “Static characterization and parameter extraction in MOS transistors”, Microelectronic engineering, Vol. 40, pp. 181-186, 1998.

[125] P. R. Karlsson and K. O. Jeppson, IEEE Trans. on Semiconductor Manufacturing, Vol. 9, pp. 215-222, 1996.

[126] H. Brut, A. Juge, and G. Ghibaudo, “New approach for the extraction of the gate voltage dependent series resistance and channel length reduction in CMOS transistors.”, Proc. of ICMTS, pp. 188-193, 1997.

[127] K. Yamaguchi, H. Asimiro, M. Yamawaki, and S. Asai, “A new variational method to determine effective channel length and series resistances of MOSFET’s”, in Proc. of ICMTS, pp. 123-126, 1997.

[128] Y. Denis, F. Monsieur, G. Ghibaudo, J. Mazurier, E. Josse, D. Rideau, C. Charbuillet, C. Tavernier, H. Jaouen, “New compact model for performance and process variability assessment in 14nm FD-SOI CMOS technology”, IEEE Proc. of ICMTS, pp. 59-64, 2015

[129] C-L. Lou, W-K. Chim, D. S-H. Chan Y. Pan, “A novel single-device DC method for extraction of the effective mobility and source drain resistances of fresh and hot carrier degraded drain engineered MOSFET’s”, IEEE Trans. Elec. Dev., Vol. 45, no. 6, 1998.

[130] J. Kim, J. Lee, I. Song, Y. Yun, J. D. Lee, B-G. Park, H. Shin, “Accurate extraction of effective channel length and source/drain resistance on ultrashort channel MOSFETs by iteration method”, IEEE Trans. Elec. Dev., Vol. 55, no. 10, 2008.

[131] Y. Denis, F. Monsieur, D. Petit, C. Tavernier, H. Jaouen, G. Ghibaudo, “A new approach for modeling drain current process variability applied to FD-SOI technology”, Proc on Ultimate Integration on

Silicon (ULIS), pp. 93-96, 2014.

[132] Y. Li and Y-Y Cho, “Intelligent BSIM4 Model Parameter Extraction for Sub-100 nm MOSFET Era”, Japanese Journal of Applied Physics, Vol. 43, pp. 1717, 2004.

[133] Q. Zhou, W. Yao, W. Wu, X. Li, Z. Zhu and G. Gildenblat, “Parameter extraction for the PSP MOSFET model by the combination of genetic and Levenberg-Marquardt algorithms”, In Proc. IEEE ICMTS, pp. 137-142, 2009

[134] R.A. Thakker, M.B. Patil, K.G. Anil, “Parameter extraction for PSP MOSFET model using hierarchical particle swarm optimization”, Engineering Applications of Artificial Intelligence, Vol. 22, no. 2, pp. 317-328, 2009.

[135] Y. Li and Y-H Tseng, “Hybrid Differential Evolution and Particle Swarm Optimization Approach to Surface-Potential-Based Model Parameter Extraction for Nanoscale MOSFETs”, Materials and Manufacturing Processes, Vol. 26, no. 3, pp. 388-397, 2011

[136] Coleman, T.F. and Y. Li, "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds," SIAM Journal on Optimization, Vol. 6, pp. 418-445, 1996.

[137] Coleman, T.F. and Y. Li, "On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds," Mathematical Programming, Vol. 67, Number 2, pp. 189-224, 1994.

[138] “Least-Squares (Model Fitting) Algorithms”. Available at: http://www.mathworks.com/help/optim/ug/least-squares-model-fitting-algorithms.html#broz0i4

[139] F. Payet, F. Bœuf, C. Ortolland, T. Skotnicki, “Nonuniform Mobility-Enhancement Techniques and Their Impact on Device Performance”, IEEE Trans. Elec. Dev., Vol. 55, No. 4, pp. 1050-1057, 2008

[140] L. Pham-Nguyen, C. Fenouillet-Beranger, G. Ghibaudo, T. Skotnicki, S. Cristoloveanu, “Mobility enhancement by CESL strain in short-channel ultrathin SOI MOSFETs”, Solid State Electron, Vol. 54, pp. 123-130, 2010.

[141] L. Francisco, M. Jimenez, “Parametric model calibration and measurement extraction for LFN using virtual instrumentation”, IEEE Proc. on Test Workshop (LATW), pp.1-6, 2013.

[142] D. B. M. Klaassen, “A Unified Mobility Model for Device Simulation—I. Model Equations and Concentration Dependence,” Solid-State Electronics, vol. 35, no. 7, pp. 953–959, 1992.

[143] C. Lombardi et al., “A Physically Based Mobility Model for Numerical Simulation of Nonplanar Devices,” IEEE Transactions on Computer-Aided Design, vol. 7, no. 11, pp. 1164–1171, 1988.

[144] M. Bidaud, H. Bono, C. Chaton, B. Dumont, V. Huard, P. Morin, L. Proencamota, R. Ranica, G. Ribes, “High-activation lase anneal for the 45 nm MOS technology platform”, Proc on IEEE International Conference on Advanced Thermal Processing of Semiconductors, 2007.

[145] K. Onishi, L. Kang, R. Choi, E. Dharmarajan, S. Gopalan, Y. Jeon, C. S. Kang, B. H. Lee, R. Nieh, J. C. Lee, “Dopant Penetration Effects on Polysilicon Gate HfO2 MOSFET’s”, Proc. On Symposium On VLSI Technology, pp131-132, 2001.

[146] “Overestimation of short-channel effects due to integrate coupling in advanced FD-SOI MOSFETs”, C. Navarro, M. Bawedin, F. Andrieu, S. Cristoloveanu, IEEE Trans. on Elec. Dev., vol. 61, no. 9, pp. 3274-3281, 2014.

[147] S. M. Sze, Physics of Semiconductor Devices, 2nd Ed. New York, USA, John Wiley & Sons, 1981.

[148] M. Haon, “Fully Depleted SOI: Achievements and Future Developments”, EUROSOI 2015.

[149] C. Galewski, J-C. Lou, W. G. Oldham, “Silicon Wafer Preparation for Low-Temperature Selective Epitaxial Growth”, IEEE Trans. on semiconductor manufacturing, Vol. 3, no. 3, pp. 93-98, 1990.

[150] X. Li, C. C. McAndrew, W. Wu, S. Chaudhry, J. Victory, G. Gildenblat, “Statistical Modeling With PSP MOSFET Model”, IEEE Trans. On Computer-aided design of integrated circuits and systems, vol. 29, no. 4, 2010.

[151] P. G. Drennan, C. C. McAndrew, “Understanding MOSFET mismatch for analog design”, IEEE J. Solid State Circuits, Vol. 38, no. 3, pp.450-456, 2003.

[152] G. Gildenblat, L. Xin, W. Weimin, W. Hailing, A. Jha, R. Van Langevelde, G.D.J. Smit, A.J. Scholten, D.B.M. Klaassen, “PSP: An Advanced Surface-Potential-Based MOSFET Model for Circuit Simulation”, IEEE Trans. on Elec. Dev., vol. 53, no. 9. 2006.

[153] C. C. McAndrew, I. Stevanovic, G. Gildenblat, “Extensions to backward propagation of variance of statistical modeling”, IEEE Design & Test of computers, Vol.27, no. 2, pp. 36-43, 2010.

[154] I. Stevanovic, C. C. McAndrew, “Quadratic backward propagation of variance for nonlinear statistical circuit modeling”, IEEE Trans. Computer-aided design of integrated circuits and systems, Vol. 28, no. 9, pp. 1428-1432, 2009.

[155] “Sentaurus TFM: TCAD for Manufacturing Solutions”, Synopsys Data Sheet. Available at: http://www.synopsys.com/Tools/TCAD/CapsuleModule/stfm_ds.pdf.

[156] H. G. Mohammadi, P-E. Gaillardon, M. Yazdani, G. De Micheli, “A fast TCAD-based methodology for variation analysis of emerging nano-devices”, IEEE Internationnal Symposium on Defect and Faults Tolerance in VLSI and Nanotechnology Systems (DFTS), pp. 83-88, 2013.

[157] K. Kakehi, H. Aikawa, T. Tadokoro, H. Eguchi, T. Hirayu, H. Yoshimura, T. Asami, K. Ishimaru, “An efficient manufacturing technique based on process compact model to reduce characteristic variation beyond process limit for 40nm node mass production”, Symposium on VLSI Technology, pp.90-91, 2011.

[158] S. Ruegsegger, A. Wagner, J. S. Freudenberg, D. S. Grimard, “Feedforward Control for Reduced Run-to-Run Variation in Microelectronics Manufacturing”, IEEE Transaction on semiconductor manufacturing, Vol. 12, no. 4, pp 493-502, 1999

[159] N. Jedidi, P. Sallagoity, A. Roussy, S. Dauzère-Pérès, J. Pinaton, “Feed-Forward Run-to-Run Control for Reduced Parametric Transistor Variation in CMOS Logic 0.13µm Technology”, ISSM 2007, pp. 320-324.

[160] G. James, D. Witten, T. Hastie, R. Tibshirani, “Resampling Methods” in *An Introduction to Statistical Learning*, 1st edition, New York, Springer, 2013, ch. 5, pp. 175-190

[161] J. Fox, “Bootstrapping regression models”, in Applied Regression Analysis and Generalized Linear Models, 3rd ed., USA: Sage, 2015, ch. 21, pp. 597-598.

[162] G. James, D. Witten, T. Hastie, R. Tibshirani, “Linear Model Selection and Regularization” in *An Introduction to Statistical Learning*, 1st edition, New York, Springer, 2013, ch. 5, pp. 203-228

[163] Stepwise regression, [web site] Available at : https://en.wikipedia.org/wiki/Stepwise_regression

[164] Matlab method: stepwisefit, [web site] Available at http://fr.mathworks.com/help/stats/stepwisefit.html

[165] N. R. Draper, H. Smith, “Applied Regression Analysis” Hoboken, NJ: Wiley-Interscience, pp. 307–312, 1998.

[166] P. L. Flom, D. L. Cassell, (2007) "Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use," NESUG 2007.

[167] C. Chatfield, (1995) "Model uncertainty, data mining and statistical inference," J. R. Statist. Soc. A 158, Part 3, pp. 419–466.

[168] A.N. Tikhonov and V.Y. Arsenin. Solutions of Ill-Posed Problems. Winston, Washington, 1977.

[169] K. Levenberg, "A Method for the Solution of Certain Non-Linear Problems in Least Squares", Quarterly of Applied Mathematics Vol. 2, pp. 164–168, 1944.

[170] D. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". SIAM Journal on Applied Mathematics, Vol. 11, no. 2, pp. 431–441, 1963.

[171] R. Tibshirani, “Regression shrinkage and selection via the LASSO”, J. R. Statist. Soc. B, Vol. 58, no. 1, pp. 267-288, 1996.

[172] Matlab method: LASSO [web site] Available at : http://fr.mathworks.com/help/stats/lasso.html

[173] H. Zou, T. Hastie, “Regularization and variable selection via the elastic net”, Journal of the Royal Statistical Society, Series B, Vol. 67, no. 2, pp. 301–320, 2005.

[174] R. J. Tibshirani, “A General Framework for Fast Stagewise Algorithms”, 2014, arXiv e-print (arXiv:1408.5801), available at: http://www.stat.cmu.edu/~ryantibs/papers/stagewise.pdf

[175] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, “Least angle resgression”, The Annals of Statistics, Vol. 32, no. 2, pp. 407–499, 2004.

[176] D. L. Donoho, “Fast Solution of \ell _{1} -Norm Minimization Problems When the Solution May Be Sparse”, IEEE. Trans. on Information Theory, Vol. 54, no. 11, pp. 4789-4812, 2008.

[177] A. E. Smith, D. W. Coit, “Penalty functions Handbook of Evolutionary Computation”, Section C 5.2. Oxford University Press and Institute of Physics Publishing, 1996.

[178] E. G. Ioannidis, C. G. Theodorou, S. Haendler, E. Josse, C. A. Dimitriadis, G.Ghibaudo, “Impact of source-drain series resistance on drain current mismatch in advanced Fully Depleted SOI n-MOSFETs”, IEEE Electron Device, To be published.

[179] L. Brieman, “Random Forests”, Machine Learning, Vol. 45, no. 1, pp. 5-32, 2001.

[180] G. Rech, T. Terasvirta, R. Tschernig, “A simple variable selection technique for nonlinear model”, Communications in Statistics - Theory and Methods, Vol. 30, no. 6, pp. 1227-1241, 2001.

[181] L. Rosas, M. Santoro, S. Mosci, A. Verri, S. Villa, “A Regularization Approach to Nonlinear Variable Selection”, Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS), pp.653-660, 2010.

[182] Z. Lva, H. Zhua, K. Yub, “Robust variable selection for nonlinear models with diverging number of parameters”, Statistics & probability letters, Vol. 91, pp. 90-97, 2014.

[183] S. Wu, H. Xue, Y. Wu, H. Wu, “Variable Selection for Sparse High-Dimensional Nonlinear Regression Models by Combining Nonnegative Garrote and Sure Independence Screening”, Statistica Sinica, Vol. 24, pp. 1365-1387, 2014.

[184] P. Radchenko, G. M. James, “Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions”, Journal of the American Statistical Association, Vol. 105, no. 492, pp. 1541-1553, 2010.

[185] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician, Vol. 46, no. 3, pp. 175-185, 1992

[186] C. Gershenson, “Artificial Neural Networks for Beginners”, available at http://arxiv.org/ftp/cs/papers/0308/0308031.pdf

[187] R. Rojas, “Neural Networks: A Systematic Introduction. Springer, Berlin, 1996.

[188] C. Cortes, V. Vapnik, "Support-vector networks", Machine Learning, Vol. 20, no. 3, p. 273, 1995

[189] H. Leeb, “Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process”, Bernoulli, Vol. 14, no. 3, pp. 661-690, 2008.

[190] R. Beran, “Confidence sets centered at C_p-estimators” Ann. Inst. Statist. Math., Vol. 48, pp. 1-15, 1996.

[191] R. Beran, “REACT scatterplot smoothers: Superefficiency through basis economy”, J. Amer. Statist. Assoc., Vol. 95, pp. 155–171, 2000.

[192] R. Beran, L. Dümbgen, “Modulation of estimators and confidence sets”, Ann. Statist., Vol. 26, pp. 1826–1856, 1998.

[193] F. andrieux, “Transistors CMOS decananometriques à canaux contraints sur silicium massif ou sur SOI. Fabrication, caracterisation et etude du transport”, Ph.D. dissertation, INPG, EEATS, Grenoble, France

[194] PricewaterhouseCoopers ©, “Evolving landscape of technology deals: Semiconductor Industry-Device deal trends”, Technology Institute, 2015, available at https://www.pwc.com/us/en/technology/publications/assets/semiconductor-industry-device-deal-trends.pdf .

[195] D. Rosso, “Global Semiconductor Industry Posts Record Sales in 2014” , semiconductor industry association, February 2, 2015

[196] “Global semiconductor sales from 1988 to 2014 (in billion U.S. dollars)”, Statistica ©, available at http://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/.

[197] Rob van der Meulen, Janessa Rivera, “Gartner Says Worldwide Semiconductor Sales Expected to Reach $358 Billion in 2015, a 5.4 Percent Increase From 2014”, Gartner, Stamford, January 14, 2015

[198] KPMG, “KPMG global semiconductor survey – cautious optimism continues”, 2014, available at: https://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/semiconductor-survey-2014.pdf

[199] P. Goos, B. Jones, *Optimal Design of Experiments: A Case Study Approach*, Wiley, 2011

[200] Yuping Wu, “Parallel hybrid evolutionary algorithm based on chaos-GA-PSO for SPICE model parameter extraction”, Proc. on Intelligent Computing and Intelligent Systems, pp. 688 – 692, 2009

[201] Yiming Li, Yen-Yu Cho, “Parallel Genetic Algorithm for SPICE Model Parameter Extraction”, Proc. on Parallel and Distributed Processing Symposium, 2006.

[202] W. Huyer, A. Neumaier , “Global Optimization by Multilevel Coordinate Search”, Journal of Global

Optimization, Vol. 14, no. 4, pp 331-355, 1999.

[203] Panos M. Pardalos, H. Edwin Romeijn, Handbook of global optimization, USA, Florida, Kluwer Academic Publishers

[204] Gui-Bo Ye, Yifei Chen, Xiaohui Xie, "Efficient variable selection in support vector machines via the alternating direction method of multipliers", Journal of Machine Learning Research, Vol. 15, pp. 832-840, 2011

[205] Ji Zhu, Hui Zou, "Variable Selection for the linear support vector machine", Studies in Computational Intelligence, Vol. 35, pp. 35–59, 2007

[206] A. Rakotomamonjy, "Variable Selection Using SVM-based Criteria", Journal of Machine Learning Research, Vol. 3, pp. 1357-1370, 2003

[207] Xiang Zhang, Yichao Wu, Lan Wang, Runze Li, "Variable selection for support vector machines in moderately high dimensions", Journal of the Royal Statistical Society: Series B, Vol. 78, no. 1, pp 53–76, 2016.

[208] R. May, G. Dandy, H. Maier, "Review of Input Variable Selection Methods for Artificial Neural Networks", in *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, Ch. 2, pp. 19-44, 2011

Appendices

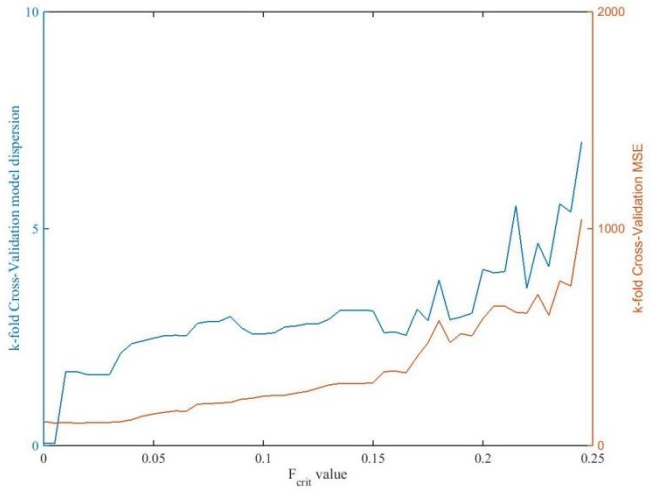
Appendix A*: Nonlinear optimization Matlab code excerpt*

The following code has been used for linear *model p*arameter extraction.

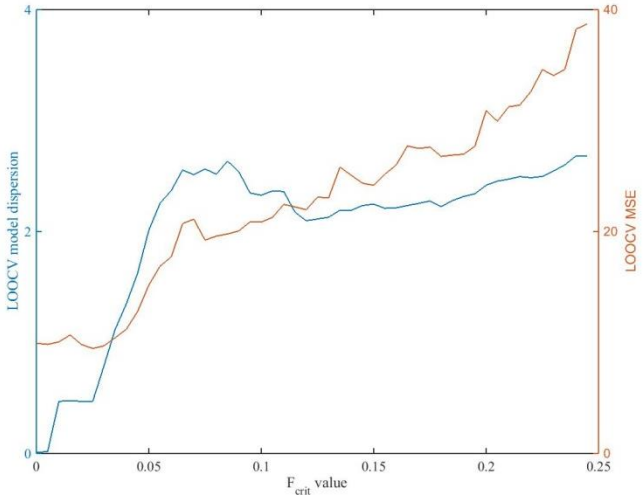
```
function guessopt =
FitIdlin(Idlin,Vg,guess,L,MSshot,Vdlin,thetaswitch,u,v,Maxit,Tol,model)

%Idlin is the linear drain current
%Idsat is the saturation drain current
%Vg is the list of linear gate voltages
%L is the list of gate length
%MSshot is the number of starting point used to find the global minima (in
the %context of multi-start problem solver algorithm)
%Vdlin is the linear drain current
%thetaswitch is a parameter that enable switching between a model with
either no %theta, theta 1 only theta 2 only or both theta 1 and theta2.
%u is the lower boundaries for model parameter to be extracted
%v is the upper boundaries for model parameter to be extracted
%Maxit is the maximum number of function evaluation allowed for extraction.
%Tol is the minimum variation of objective function and model parameter
step size %tolerated. If variation of objective function and model
parameter step size go %beyond this threshold, the optimizer stops,
considering a local minimum reached.
%model enable switching from different models (no sigma, Vtsp=Vtlin, no Lc
or all %parameters included).
% guess is the initial guess for linear model parameters

%Below, all combination of L and Vg are stacked in a matrix. The matrix is
called %"data".
ind=1;

xx=zeros(1,size(Vg,1)*numel(L));
yy=zeros(1,size(Vg,1)*numel(L));
zz=zeros(1,size(Vg,1)*numel(L));
for i=1:size(Vg,1)
    for j=1:numel(L)
        xx(ind)=Vg(i,j);
        yy(ind)=L(j);
        zz(ind)=Idlin(i,j);
        ind=ind+1;
    end
end
data=[xx;yy];

%Below options are set for the extraction procedure
opts=optimset('lsqcurvefit');
opts.TolX=Tol;
opts.TolFun=Tol;
opts.MaxIter=Maxit;
opts.MaxFunEvals=Maxit;
opts.Display='off';

%Mutlistart object is created below
ms=MultiStart;
ms.UseParallel='always';
```

```
%Following line launch the nonlinear optimizer. The function to be
optimized is %Rlinfun that takes (x,data,L,thetaswitch,model,Vdlin) as
parameters where x is %the model parameter guess (it changes from one
iteration to another).
%xdata is the matrix of Vg-L combination called “data”.
%ydata are the measurements
%lower and upper bound are specified via parameter 'lb' and 'ub'
%'options' specify the options chosen before and stored in the object
called %“opts”.

problem=createOptimProblem('lsqcurvefit','objective',@(x,data)Rlinfun(x,dat
a,L,thetaswitch,model,Vdlin),'xdata',data,'ydata',Vdlin./zz,'x0',guess,'lb'
,u,'ub',v,'options',opts);
guessopt=run(ms, problem, MSshot);
end
```

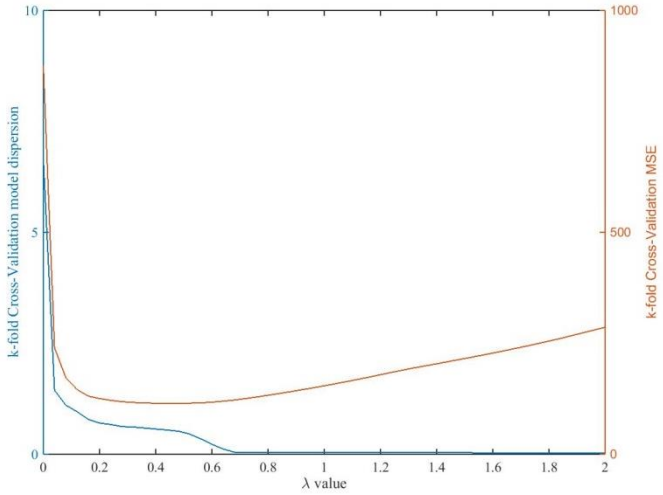
Appendix B*: K-fold CV and LOOCV plots for model parameter PCM building*



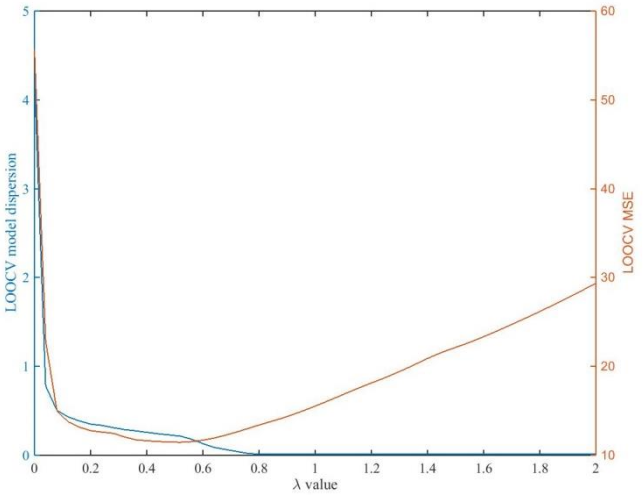
(a.1)



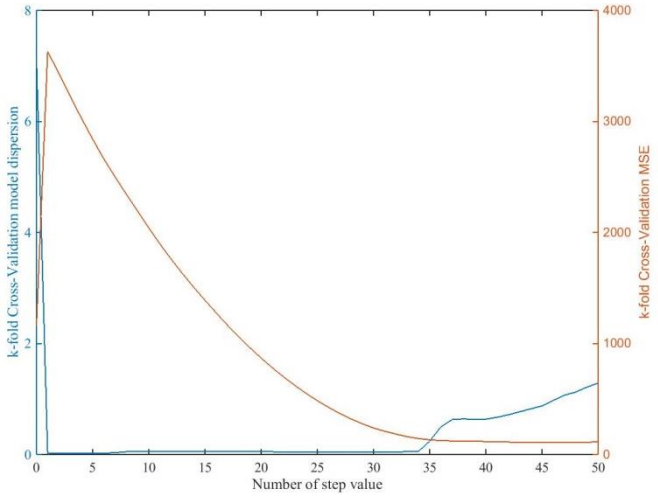
(a.2)



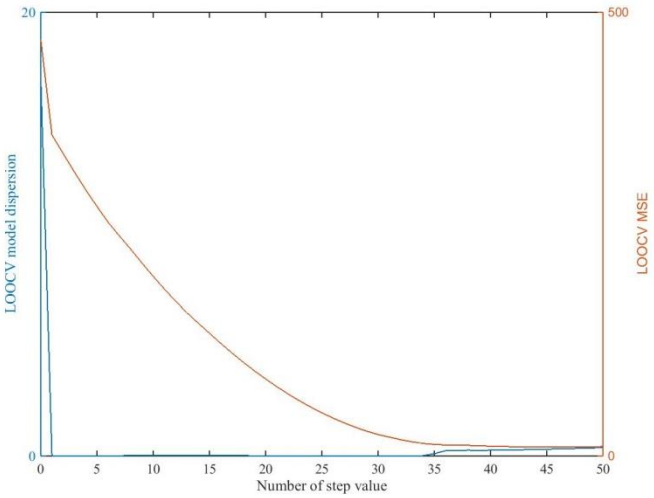
(b.1)



(b.2)

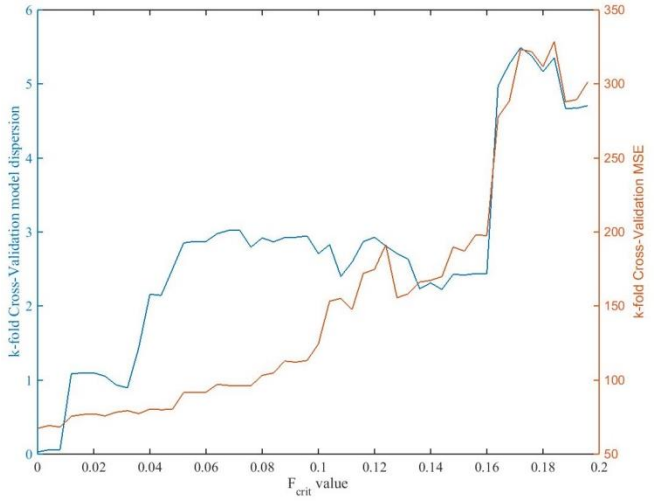


(c.1)

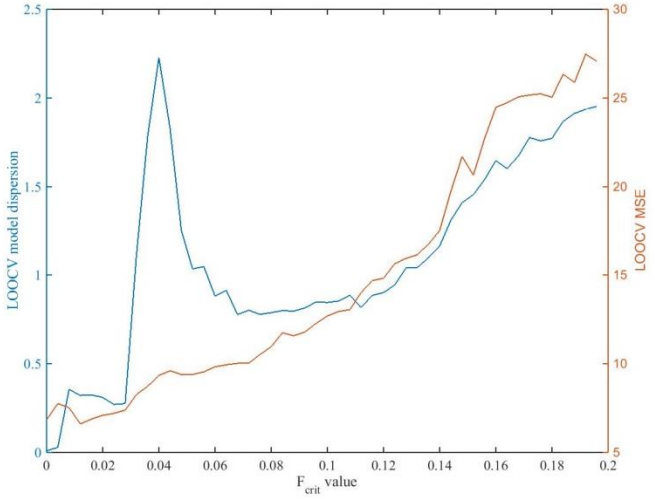


(c.2)

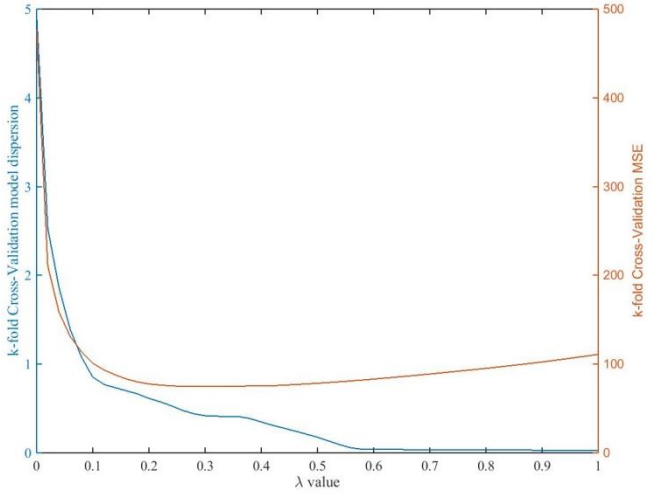
Figure 0-1: K-fold CV (1) and LOOCV (2) MSE and model dispersion test for R_0 PCM against the calibration parameter of stepwise regression (a), LASSO (b) and LARS (c) method.



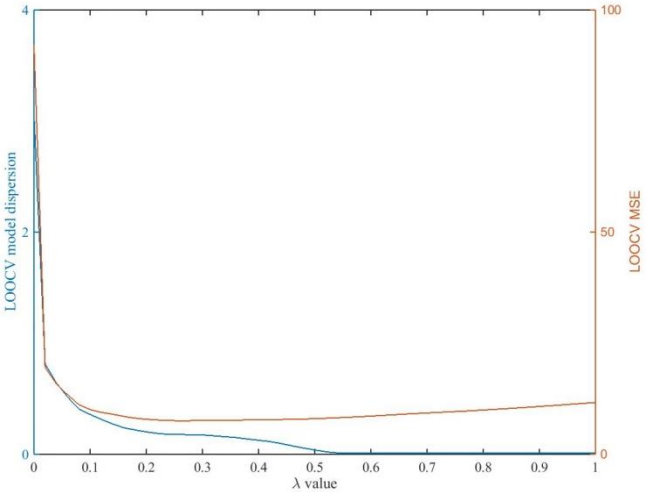
(a.1)



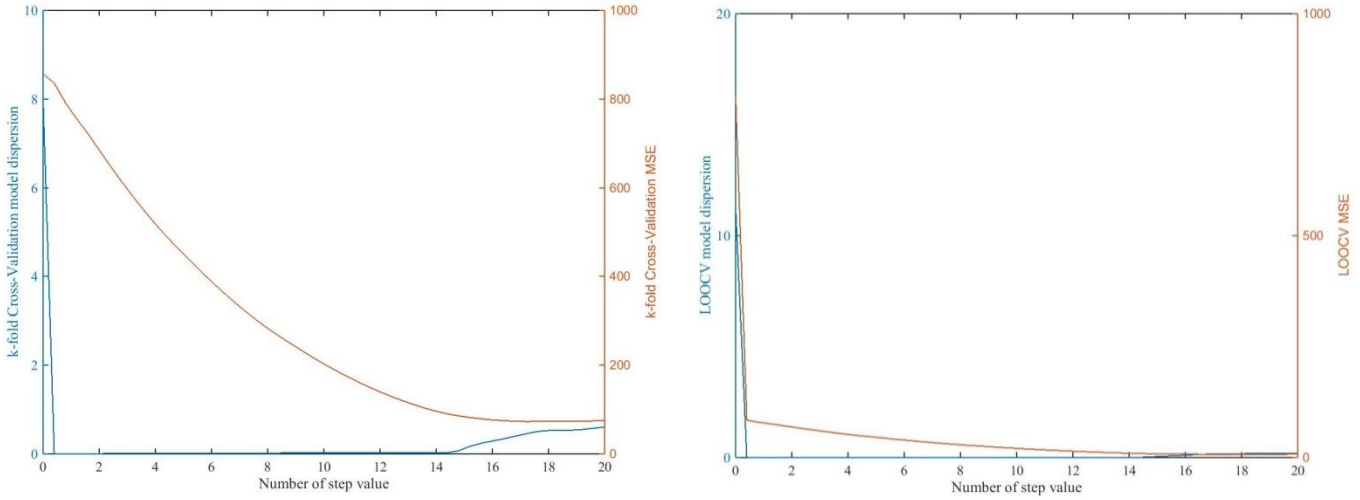
(a.2)



(b.1)

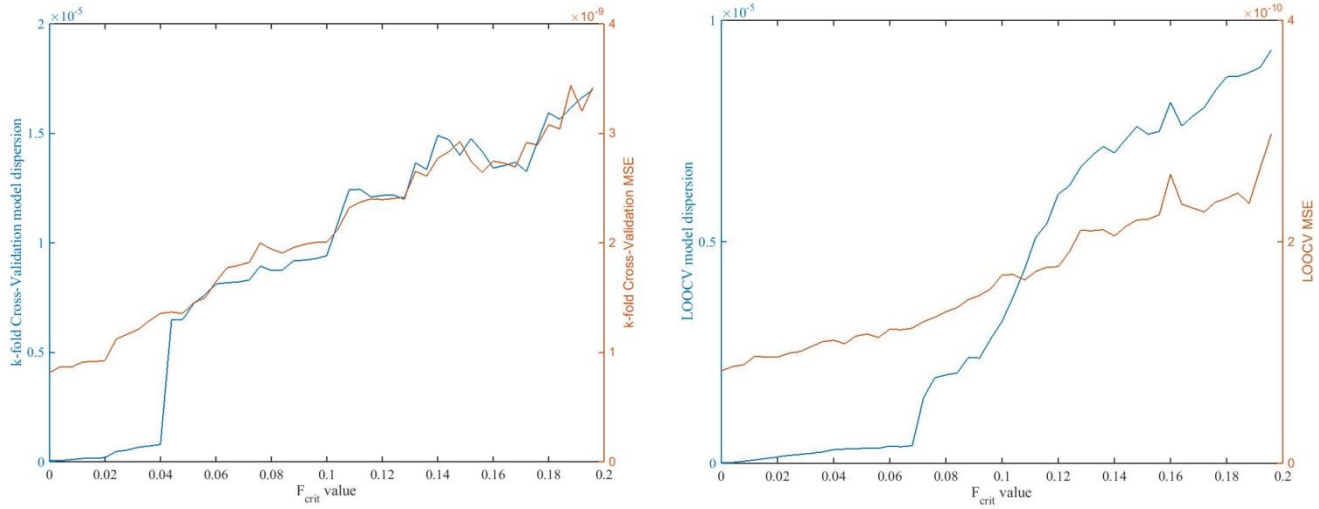


(b.2)

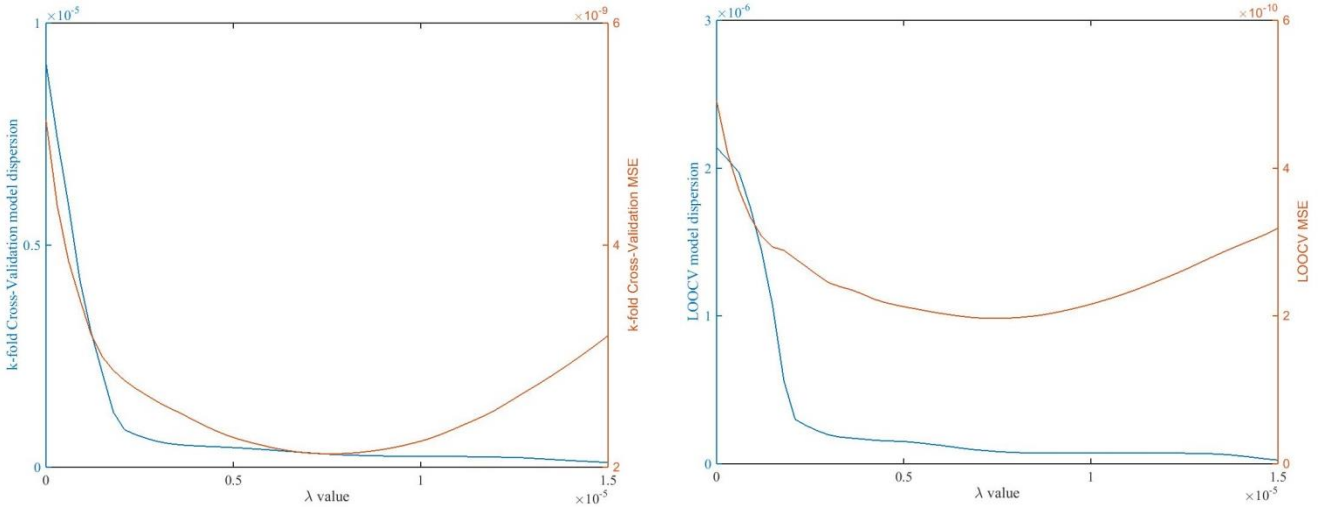


(c.1) (c.2)

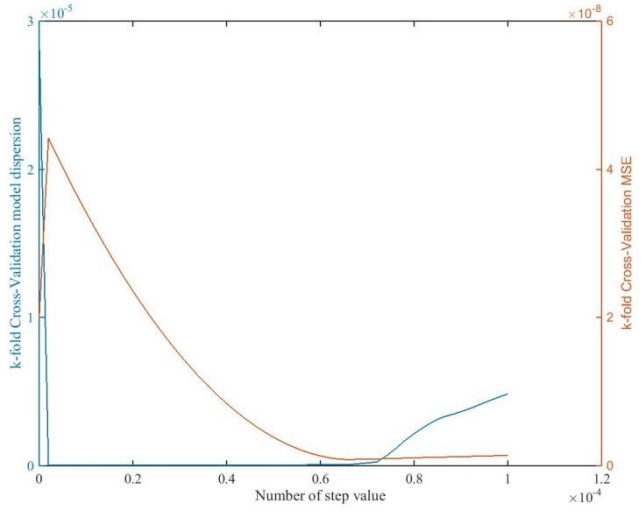
Figure 0-2: K-fold CV (1) and LOOCV (2) MSE and model dispersion test for σ PCM against the calibration parameter of stepwise regression (a), LASSO (b) and LARS (c) method.



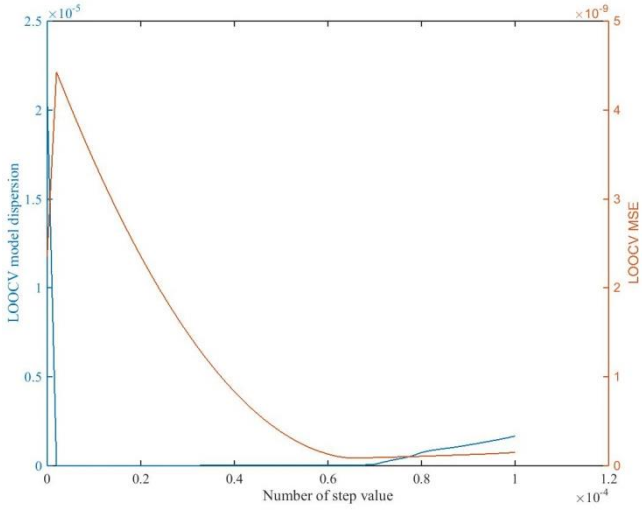
(a.1) (a.2)



(b.1) (b.2)

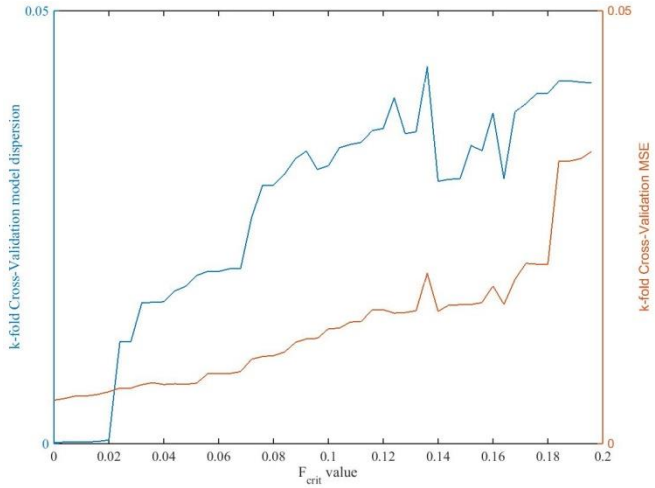


(c.1)

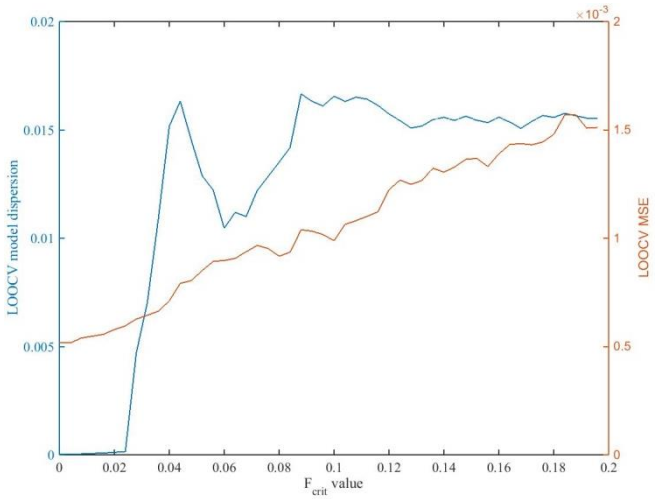


(c.2)

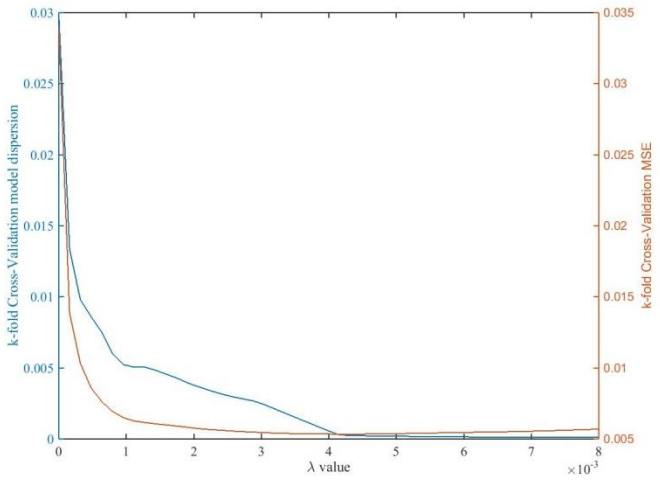
Figure 0-3: K-fold CV (1) and LOOCV (2) MSE and model dispersion test for $\mu_0.C_{ox}$ PCM against the calibration parameter of stepwise regression (a), LASSO (b) and LARS (c) method.



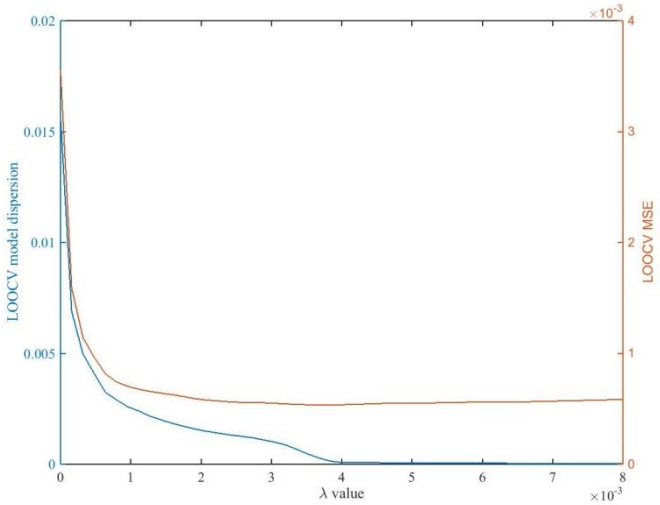
(a.1)



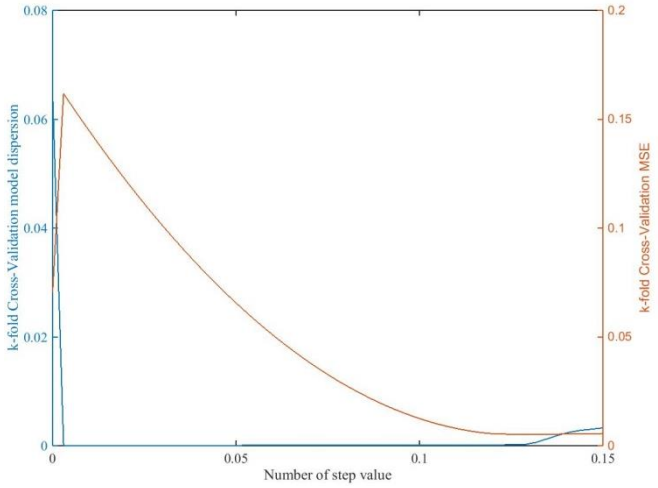
(a.2)



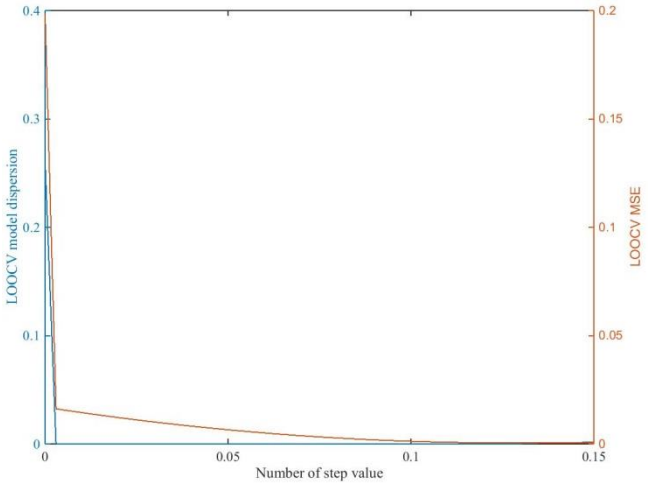
(b.1)



(b.2)

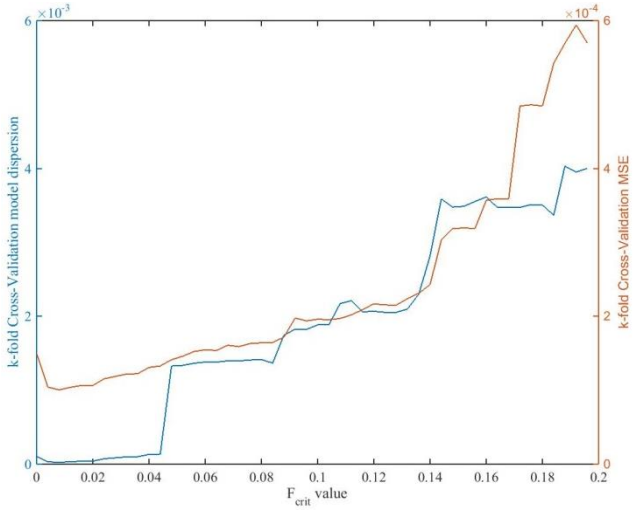


(c.1)

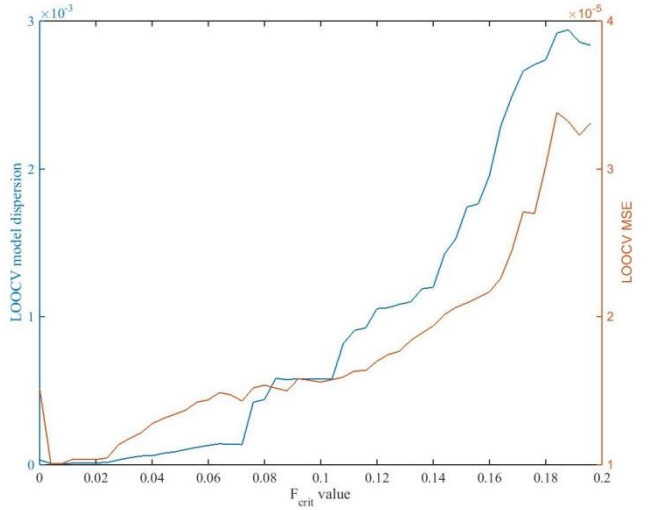


(c.2)

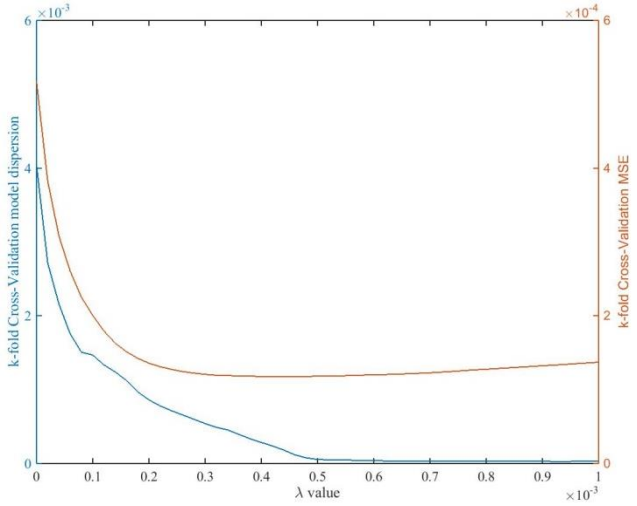
Figure 0-4: K-fold CV (1) and LOOCV (2) MSE and model dispersion test for θ_2 PCM against the calibration parameter of stepwise regression (a), LASSO (b) and LARS (c) method.



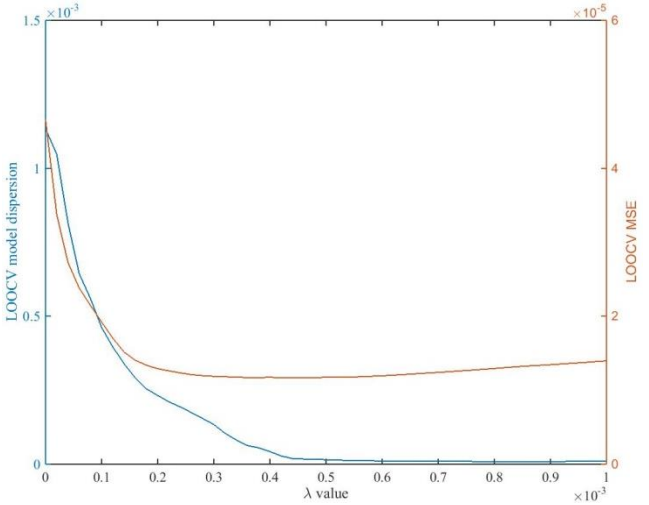
(a.1)



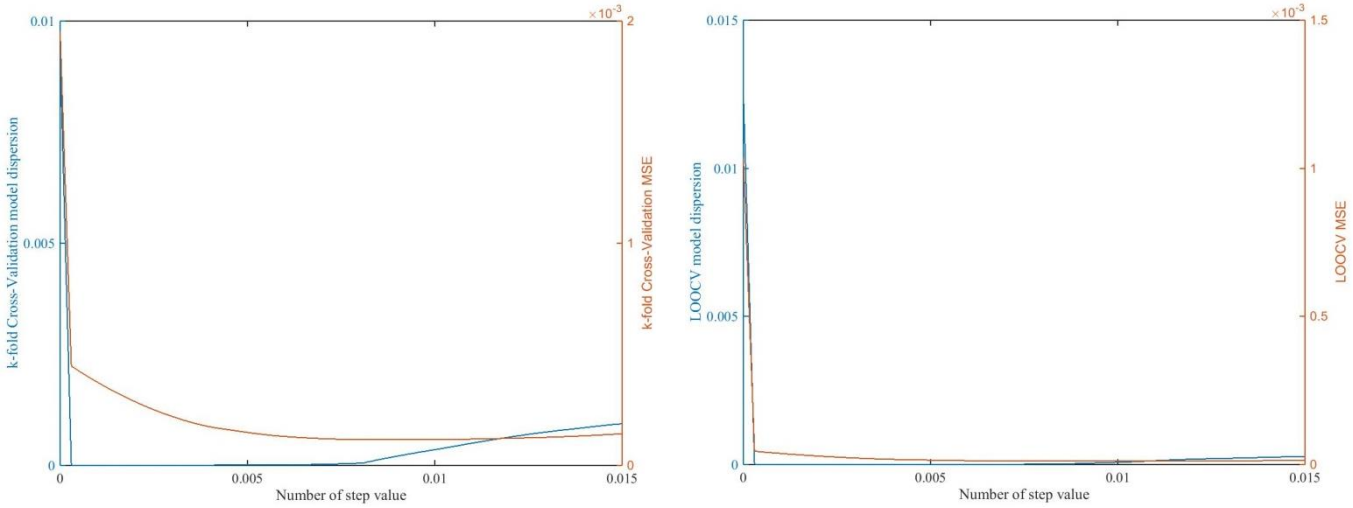
(a.2)



(b.1)

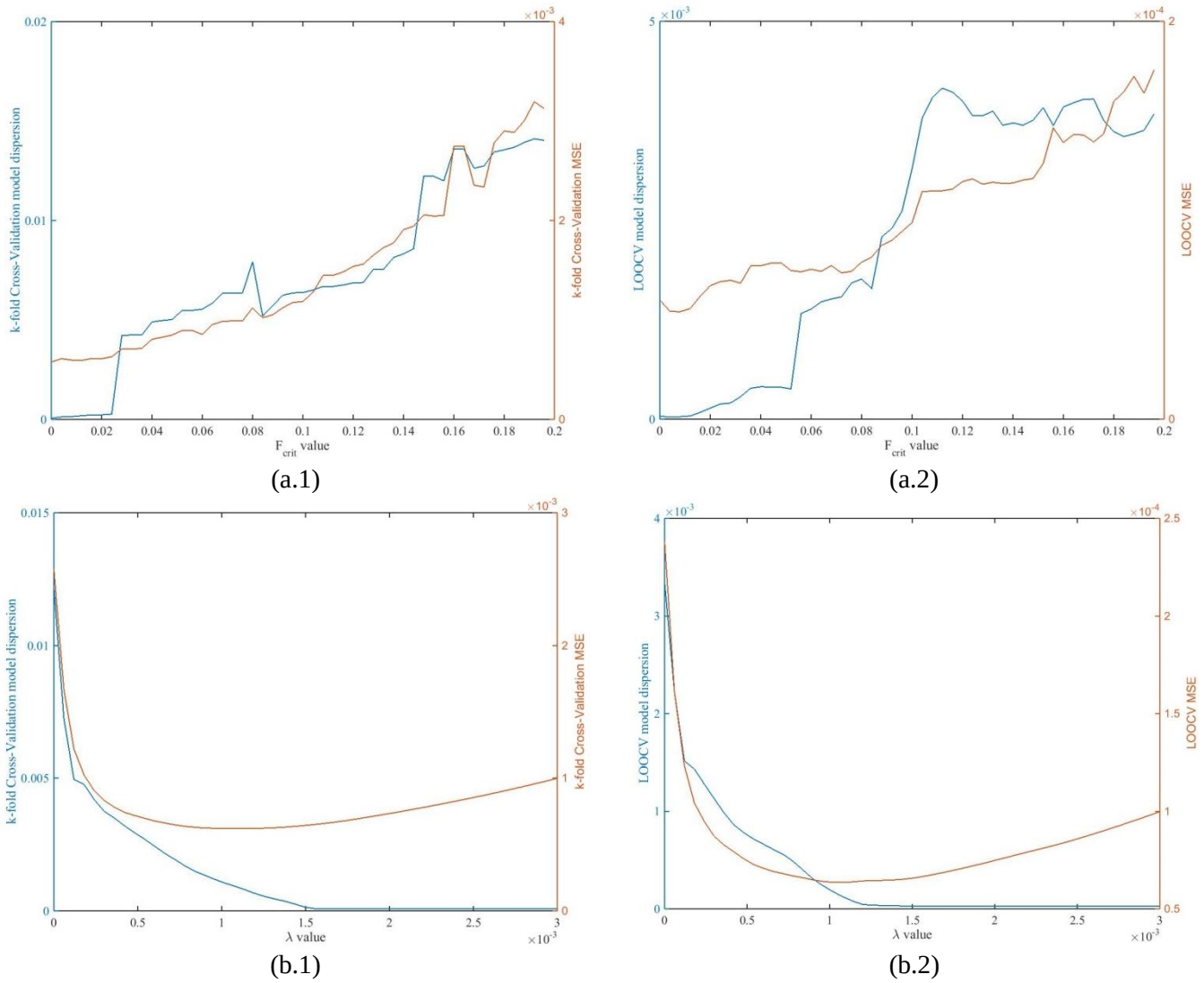


(b.2)



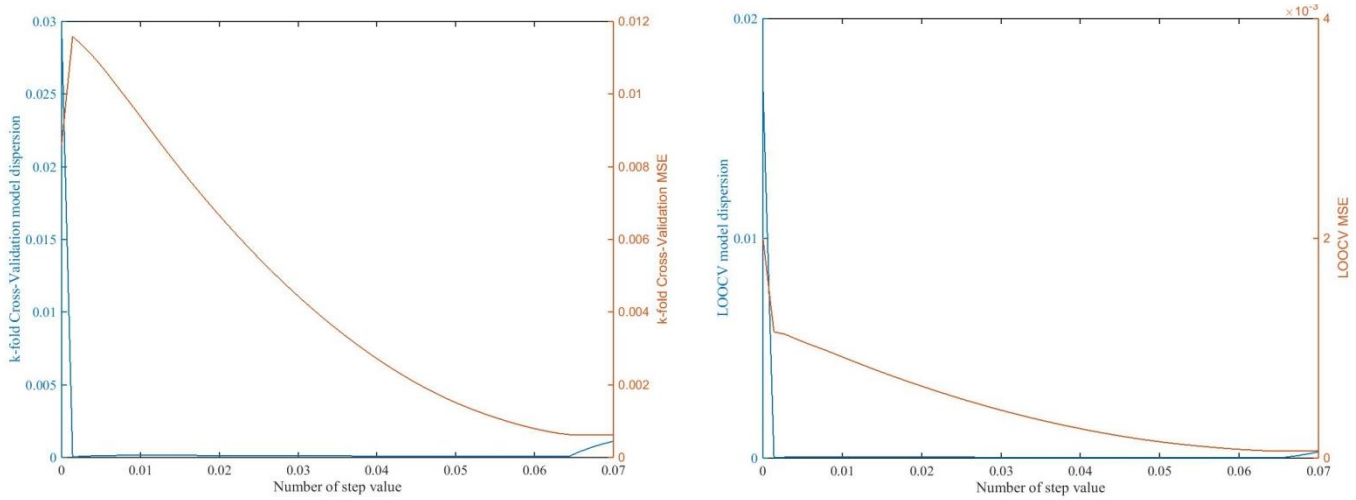
(c.1) (c.2)

Figure 0-5: K-fold CV (1) and LOOCV (2) MSE and model dispersion test for V_{tlin} PCM against the calibration parameter of stepwise regression (a), LASSO (b) and LARS (c) method.



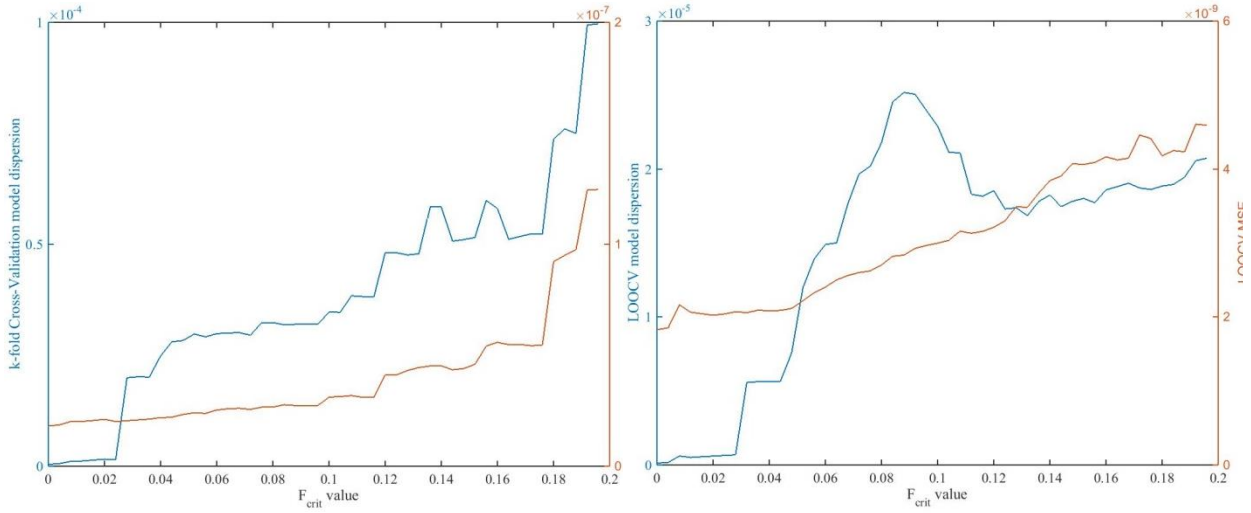
(a.1) (a.2)

(b.1) (b.2)

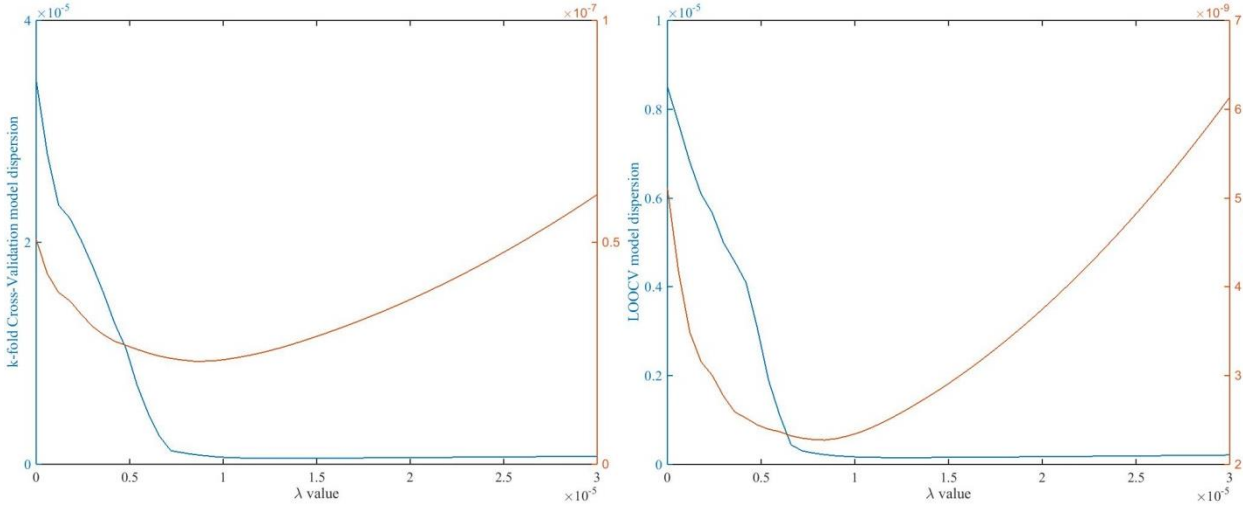


(c.1) (c.2)

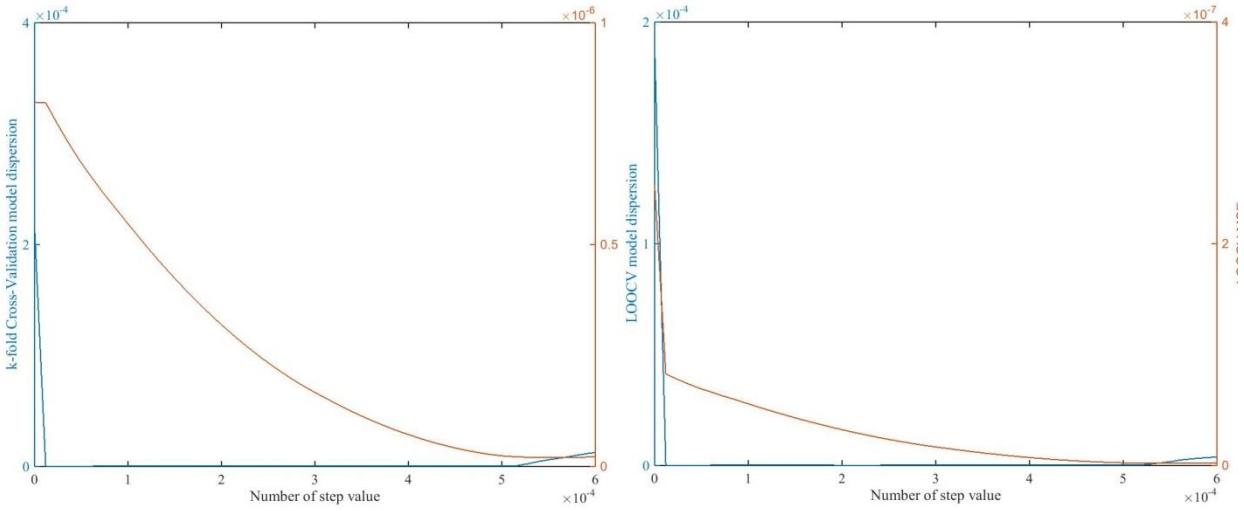
Figure 0-6: K-fold CV (1) and LOOCV (2) MSE and model dispersion test for V_{tsat} PCM against the calibration parameter of stepwise regression (a), LASSO (b) and LARS (c) method.



(a.1) (a.2)



(b.1) (b.2)



(c.1) (c.2)

Figure 0-7: K-fold CV (1) and LOOCV (2) MSE and model dispersion test for $v^*.C_{ox}$ PCM against the calibration parameter of stepwise regression (a), LASSO (b) and LARS (c) method.

Appendix C*: List of publications*

1st author:

Y. Denis, F. Monsieur, G. Ghibaudo, J. Mazurier, E. Josse, D. Rideau, C. Charbuillet, C. Tavernier, H. Jaouen, "New compact model for performance and process variability assessment in 14nm FD-SOI CMOS technology", IEEE Transaction on manufacturing, *Currently reviewing*.

Y. Denis, F. Monsieur, G. Ghibaudo, J. Mazurier, E. Josse, D. Rideau, C. Charbuillet, C. Tavernier, H. Jaouen, "New compact model for performance and process variability assessment in 14nm FD-SOI CMOS technology", IEEE Proc. of ICMTS, pp. 59-64, 2015

Y. Denis, F. Monsieur, D. Petit, C. Tavernier, H. Jaouen, G. Ghibaudo, "A new approach for modeling drain current process variability applied to FD-SOI technology", Proc on Ultimate Integration on Silicon (ULIS), pp. 93-96, 2014.

2nd author:

F. Monsieur, Y. Denis, D. Rideau, J. Lacord, V. Quenette, G. Gouget, C. Tavernier, H. Jaouen, 'The importance of the spacer region to explain short channels mobility collapse in 28 nm Bulk and FD-SOI technologies', IEEE Proc. on ESSDERC 2014.

Appendix D*: Résumé en Français*

La vitesse de succession des nœuds technologiques s'est ralentie récemment [193] à cause des nouveaux défis technologiques rencontrés. Parmi ces obstacles, on trouve la part croissante de la variabilité du procédé et local stochastique, due à une complexité croissance du processus de fabrication et à la miniaturisation, en plus de la difficulté à réduire la longueur du canal. Certains de ces obstacles requièrent l'adoption de nouvelles architectures, différente de celle traditionnelle (Bulk). Cependant, ces nouvelles architectures requièrent de plus lourds investissements pour être industrialisées. L'augmentation de la complexité et du temps de développement induit une augmentation des investissements financiers. Bien que le marché du semi-conducteur soit large et que les ventes augmentent continument, comme montré dans la, la croissance de l'industrie donne des tendances moins convaincantes, fortement dépendantes de la conjecture économique [194].

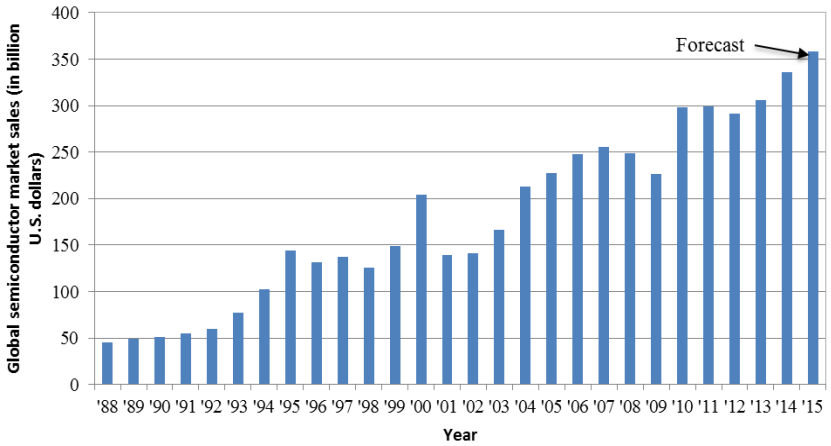


Figure 8 : Chiffre d'affaire du marcher du semi-conducteur international par an depuis 1988 à 2015 [195][196][197].

Par conséquent, les marges d'investissement diminuent et le secteur de la R&D (qui requière de larges financements avec une durée de rentabilisation de l'investissement long) devra faire face à ce problème. Une étude récente a demandé aux chefs d'entreprise du secteur du semi-conducteur du monde entier, quel sera le plus gros problème auquel l'industrie devra faire face pendant les 3 prochaines années. La première réponse est l'augmentation des coûts de R&D, suivit des ruptures technologiques puis le prix élevé des nouvelles installations et équipements [198]. De fait il y a un réel besoin d'améliorer le développement et l'optimisation de la fabrication des dispositifs. Ce travail donne quelques pistes pour atteindre ce but. Dans le but de développer et optimiser un dispositif, les ingénieurs font largement appel à des essais successifs et à l'expertise scientifique. Cette approche est apparu être la plus efficace et fiable jusqu'à maintenant. Cependant, avec l'accroissement de la complexité des nouvelles architectures et de la variabilité stochastique, cette approche demande de plus en plus d'essais, accroissant dramatiquement le coût de développement. Dans le but de résoudre ce problème, l'idée est de minimiser le nombre d'essais nécessaire pour obtenir le du processus de fabrication optimal. Le du processus de fabrication optimal est celui qui permet d'obtenir le dispositif dont les performances électriques et leur dispersion atteignent les objectifs prédéterminés.

Un moyen de trouver le du procédé optimal, sans faire appel à de nombreux essaies sur silicium, est d'utiliser des modèles compacts couplé à l'outil de simulation TCAD. En effet, un model précis et rapide à calculer qui établit les relations entre les paramètres du procédé et électrique peut être utilisé en entré d'un algorithme d'optimisation qui, à son tour, peut trouver un du procédé optimal. L'outil TCAD fait appel à une calibration physique précise et fiable. Cependant les simulations sont longues (de l'ordre de quelques heures). Dans la mesure où la plupart des algorithmes de simulation requièrent un grand nombre d'évaluations (plus de 1000), faire uniquement appel à la TCAD ne pourra pas donner de résultats dans un temps raisonnable. De plus, la calibration d'une simulation TCA est une tâche difficile qui requière de larges études physiques. Au contraire, les modèles compacts (comme

BSIM, PSP, …) sont rapide à calculer et ne requière qu'une caractérisation électrique complète du dispositif pour être calibré. Cependant la plupart des paramètres modèles ne sont pas directement reliés à un paramètre du procédé, c'est-à-dire, quelques paramètres ont une interprétation physique complexe. Par exemple, l'interprétation physique de la longueur de canal effective (qui est largement utilisé dans les modèles compacts) a été le sujet de nombreuses investigations [**50**][**83**][**91**][99]-[101][106][114][127][130]. De fait, utiliser un modèle compact pour optimiser les performances électriques ne pourra conduire à un du procédé optimal.

L'idée qui a été développé dans cette thèse est de combiner la TCAD avec un modèle compact dans le but de construire et calibrer ce qu'on appelle un Du procédé Compact Model (PCM). Un PCM est modèle analytique qui lie les paramètres du procédé et électriques du MOSFET. Il tire les bénéfices à la fois de la TCAD (dans la mesure où il lie les paramètres électriques aux paramètres du procédé) et du modèle compact (puisque le modèle est analytique et rapide à calculer). Notre PCM est décomposé en étages. En commençant par les paramètres du procédé, le premier étage est formé de plusieurs polynômes qui relient les paramètres du procédé avec les paramètres modèles d'un modèle compact typiquement basé sur la tension de seuil. Le second étage est le modèle compact, qui donne les paramètres électrique en sortie. Un schéma d'entrée/sortie du PCM est présenté dans la Figure 9.

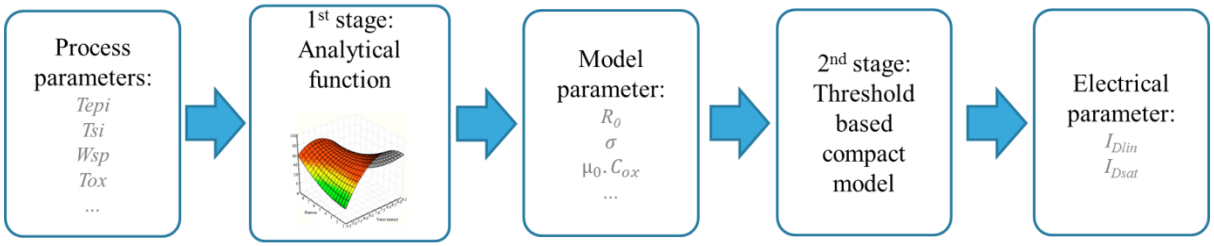


Figure 9: Schéma du PCM à deux phases

6.3 Résumé de la thèse

Après une brève introduction, dans le chapitre 2, ce manuscrit détail le modèle compact utilisé pour le deuxième étage du PCM qui sépare le courant linéaire et saturé en paramètres modèle tels que la résistance d'accès, la mobilité des porteurs, la tension de seuil… La démonstration issue des bases de la physique du semi-conducteur a été développée. Dans un premier temps, la capacité MOS a été considérée pour déterminer la formule de la charge d'inversion de même que la tension de seuil pour le cas du MOS bulk. Cette équation a été adaptée au cas du dispositif UTBB (Ultra thin Body and Box). L'effet de la concentration de dopant dans le canal, de l'épaisseur réduite du canal et du BOx (Buried Oxide) sur la tension de seuil et la charge d'inversion ont été traité. Un modèle compact pour la mobilité a été proposé, où la rugosité de surface, l'interaction avec les charges Coulombienne et les phonons ainsi que les défauts neutres, le transport balistique, la vitesse de saturation et d'injection des porteurs ont été pris en compte. Ensuite le courant de drain linéaire et en saturation ont été introduit, basée sur les formules de mobilité, de tension de seuil et de charge d'inversion proposée précédemment. Dans les dispositifs réels, les modèles compacts doivent prendre en compte la résistance d'accès. De fait, cet aspect a été traité et les modèles compacts analytiques pour les régimes linéaires (I_{Dlin}) et de saturation (I_{Dsat}) ont été adapté. Les équations correspondantes sont les suivantes :

$$Id_{lin} = \frac{V_{DS}}{R_{tot}} \tag{161}$$

Où la résistance totale normalize par la largeure du transistor (notée R_{tot}) est donné par l'équation suivante:

$$R_{tot} = R_0 + \frac{\sigma}{V_{GS} - V_{tLDR}} + \frac{W\left(1+\frac{L_c}{L}\right)}{\beta}\left(\frac{1}{\left(V_G - V_t - \frac{V_{DS}}{2}\right)} + \theta_1 + \theta_2\left(V_G - V_t - \frac{V_{DS}}{2}\right)\right) \quad (162)$$

La résistance total est simplement la somme des résistances d'accès constantes (noté R_0), des résistances d'accès dépendant de V_G (représenté par le terme σ) et de la résistance canal.

L'expression du courant de drain en saturation est donnée comme suit:

$$Id_{sat} = \frac{Id'_{sat}}{1 + G_m \cdot R_S} \quad (163)$$

Où $R_s = \frac{R_0 + \frac{\sigma}{V_{GS} - Vt_{LDR}}}{2}$ et Id'_{sat} est le courant de drain intrinsèque en saturation:

$$Id'_{sat} = \frac{W}{L + L_c}\mu_{eff} C_{ox}\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)V_{Dsat} \quad (164)$$

G_m est la dérivée du courant de drain par rapport à V_G et s'exprime comme suit :

$$G_m = 4\frac{W}{L + L_c}\mu_{eff} C_{ox} V_{Dsat} \cdot \left(A - \theta_2\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)^2\right) \quad (165)$$

Dans cette expression, $A = 1 + \frac{1}{L}\left(\frac{V_{Dsat}\cdot\mu_0}{v^*}\right)$ et μ_{eff} est la mobilité effective prenant compte des mécanismes d'interaction avec les porteurs (rugosité de surface, interaction Coulombiennes et interaction avec les phonons), la vitesse de saturation des porteurs ainsi que le transport balistique:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta_1\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right) + \theta_2\left(V_{GS} - V_t - \frac{V_{Dsat}}{2}\right)^2 + \frac{V_{Dsat}\cdot\mu_0}{L v^*}} \quad (166)$$

De même, V_{Dsat} est la tension de drain en saturation et est déduit comme étant la tension V_{DS} tel que $\frac{dId_{lin0}}{dV_{DS}} = 0$ où Id_{lin0} est le courant de drain linéaire intrinsèque. L'expression de V_{Dsat} est la suivante:

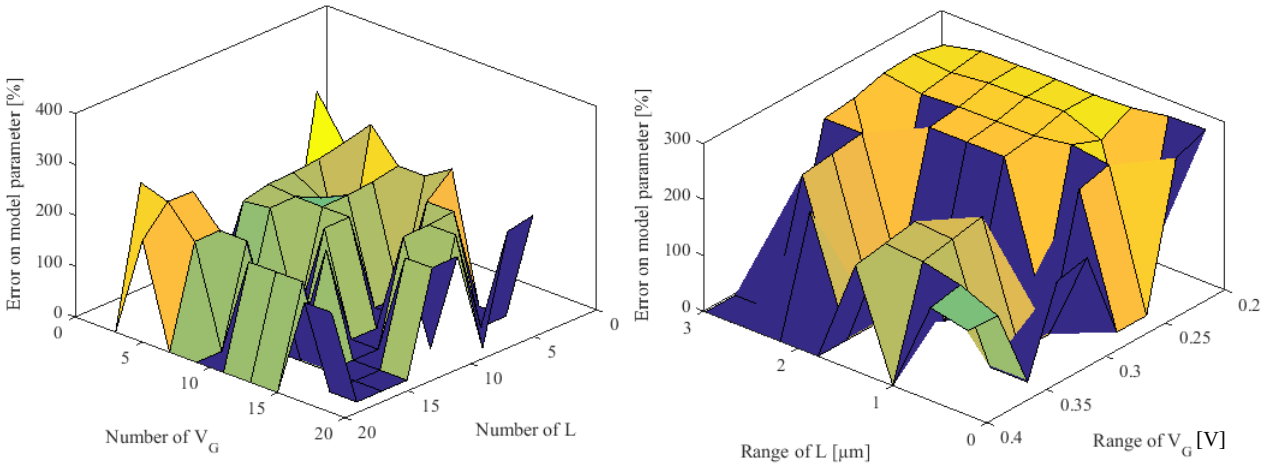
$$V_{Dsat} = 2\frac{u - \sqrt{u.\left(1 + 2.\frac{\mu_0}{v^*.L}(V_G - V_{tsat})\right)}}{\theta_1 - \frac{2\mu_0}{L.v^*} + (V_G - V_{tsat})\theta_2} \quad (167)$$

Où $u = 1 + \theta_1(V_G - V_{tsat}) + \theta_2(V_G - V_{tsat})^2$. Ces formulations sont valides en régime de forte inversion. L'avantage de ces formulation est qu'elles sont analytiques et donc rapide à calculer. Ces atouts sont de rigueur pour permettre une application industrielle et pour extraire les paramètres sur une grande quantité de dispositifs mesurés avec un échantillon de tension très réduit. Cependant des simplifications ont été nécessaires pour atteindre ces objectifs. Il a été montré que l'impact de ces simplifications est acceptable en comparaison avec des calculs numériques rigoureux.

Une large partie de cette thèse est dévouée à la procédure de construction et de calibration de ce PCM. Cela peut être fait avec l'aide de la TCAD et/ou des mesures silicium. La procédure requière quelques

valeurs de courant de drain mesurées ou simulées sur des transistors de différentes longueurs, à des tensions de grille différentes. Sur la base de ses données, les paramètres modèles sont accessibles via une procédure d'extraction spécifique développée dans le chapitre 3. Cette méthode fait appel à un nombre limité de mesures, ce qui rend possible le contrôle continu et exhaustif de la ligne de production. La méthode est décomposée en en trois étapes. La première étape consiste à extraire les paramètres du modèle linéaire via un ajustement des moindres carrés linéaire. Ensuite, ces valeurs sont utilisées comme une première estimation en entrée d'un optimiseur non linéaire qui affine les valeurs des paramètres. Enfin, les paramètres du modèle de saturation sont extraits via un ajustement des moindres carrés non linéaire.

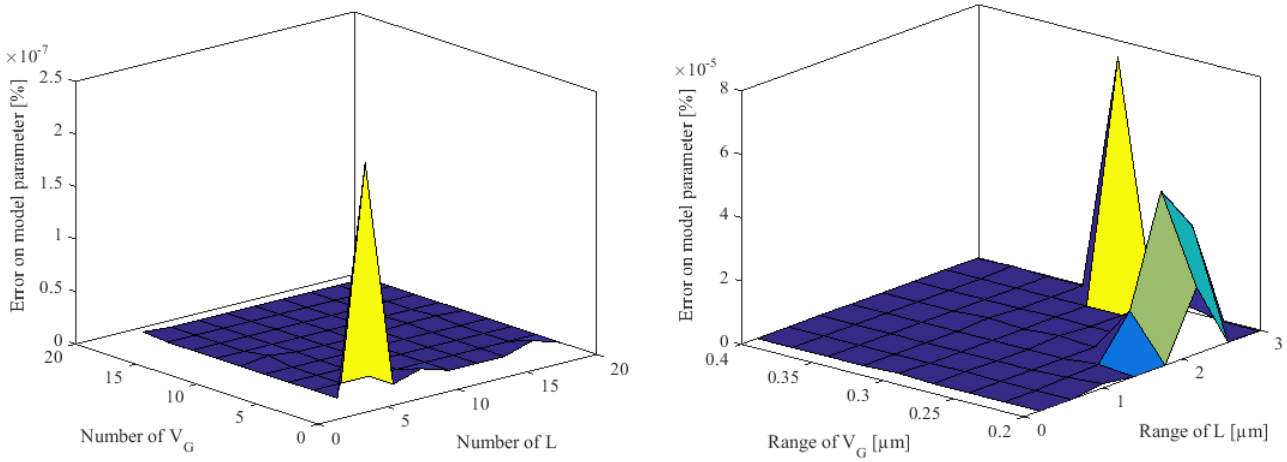
Cette méthode a été testée pour évaluer sa robustesse. Des tests ont été effectués avec des données synthétiques pour valider la taille et l'étendue de l'échantillon de mesure. Nous avons vu que l'étendue et la taille de l'échantillon de données disponibles dans les mesures silicium sont trop petites pour extraire correctement tous les paramètres du modèle. Les résultats sont donnés dans la Figure 3-10.



(a) (b)

Figure 10: Erreur sur les paramètres du modèle extrait à partir de données synthétiques avec différentes taille (a) et étendue (b) d'échantillon. La totalité des paramètres ont été considéré pour l'extraction.

En supprimant successivement chacun des paramètres modèle a permis de montrer que θ_1, $Vt_{\text{région faiblement dopée}}$ et L_c sont les paramètres modèle les moins significatifs. Les tests d'extraction ont été exécutés de nouveau dans les cas où certain paramètres modèle ont été retirés. Il a été montré que dès lors qu'un paramètre est supprimé, l'extraction fonctionne très bien. Ainsi, la suppression d'un paramètre permet une extraction solide avec un biais minimal dans le modèle.



(a) (b)

Figure 11: Erreur sur les paramètres du modèle extrait à partir de données synthétiques avec différentes taille (a) et étendue (b) d'échantillon. Lc n'est pas extrait ici (Lc=0).

À la suite de cette étude, l'effet du bruit de mesure dans le procédé d'extraction a été étudié. Il a révélé qu'une petite quantité de bruit peut conduire à de fortes erreurs dans l'extraction du modèle. Une étude TCAD du modèle compact de mobilité a montré que l'utilisation simultanée de θ_1 et θ_2 dans le modèle peut conduire à une forte incertitude sur les résultats de l'extraction. Retirer θ_1 rend les extractions plus robustes face au bruit. En fixant θ_1 à 0, des tests de bruit ont été effectuées sur la base de données synthétiques, construites à partir des paramètres modèle extraits sur des caractéristiques I_D-V_G en forte inversion, mesurées sur les dispositifs nMOS et pMOS des nœuds technologique 28 et 14 nm FD-SOI. Les résultats ont montré un niveau raisonnable de bruit dans les paramètres modèle extraits, avec 1% du bruit sur les paramètres électriques comme le montre la Figure 3-19.

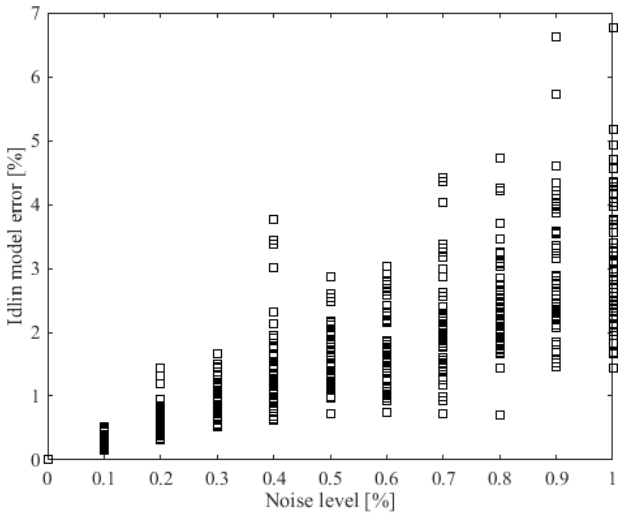


Figure 12: RMS error against artificial noise induced in the synthetized data.

Ces tests ont montré qu'une attention particulière doit être portée au modèle utilisé pour l'extraction. Nous suggérons d'abord la mise en θ_1à 0 afin de réduire l'impact du bruit dans les mesures. Ensuite, en fonction de l'appareil, un ou deux paramètres doivent être enlevés (Vt_{LDR} et / ou L_c). Afin de vérifier la validité de ces simplifications, les résultats d'extraction doivent être vérifiés. Pour les simulations TCAD, la cohérence physique entre variation de paramètre modèle et du procédé a été vérifiée. Quant aux extractions sur données silicium, des courbe de corrélation ont été faites montrant que les paramètres du modèle sont la plupart du temps décorrélées. L'absence de corrélation entre paramètre modèle assure la robustesse de l'extraction et permet de tirer des conclusions sur l'impact de la variation des paramètres modèle sur le courant de drain.

La procédure d'extraction a été exécutée sur un plan d'expériences simulé en TCAD. Le plan d'expériences prend en compte différents paramètres du procédé (résistance externe, épaisseur d'épitaxie, épaisseur du SOI, largeur des éspaceurs, la dose implantée, la température de recuit, l'épaisseur de la couche isolante, épaisseur de diélectrique à haute permittivité). Ce plan d'expérience est donnée dans le Table 5-4 où les valeurs -1 0 et 1 correspondent à celles données dans le Table 5-5 et

Experiment\Parameter	Tepi	Wsp	Tsi	T_{il}	Fdose	Tspike	Rext	Qhk
1(reference)	0	0	0	0	0	0	0	0
2	-1	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0
4	0	-1	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0

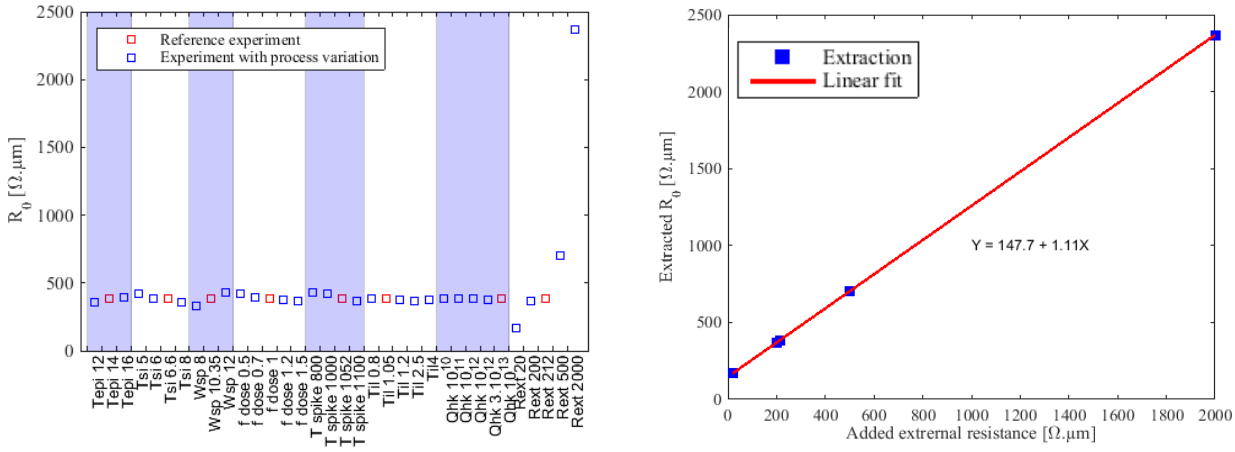
6	0	0	-1	0	0	0	0	0
7	0	0	1	0	0	0	0	0
8	0	0	0	-1	0	0	0	0
9	0	0	0	1	0	0	0	0
10	0	0	0	0	-1	0	0	0
11	0	0	0	0	1	0	0	0
12	0	0	0	0	0	-1	0	0
13	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	-1	0
15	0	0	0	0	0	0	1	0
16	0	0	0	0	0	0	0	-1
17	0	0	0	0	0	0	0	1

Tableau 1: Plan d'expérience composite face centrée simulé en TCAD

Variable Level	Tepi	Wsp	Tsi	Til	fdose	Tspike	Qhk	Rext
-1	12	8	5	0.8	0.5	900	10^{10}	70
0	14	10	6.5	1	1	1000	10^{12}	100
1	16	12	8	2	1.5	1100	10^{13}	130

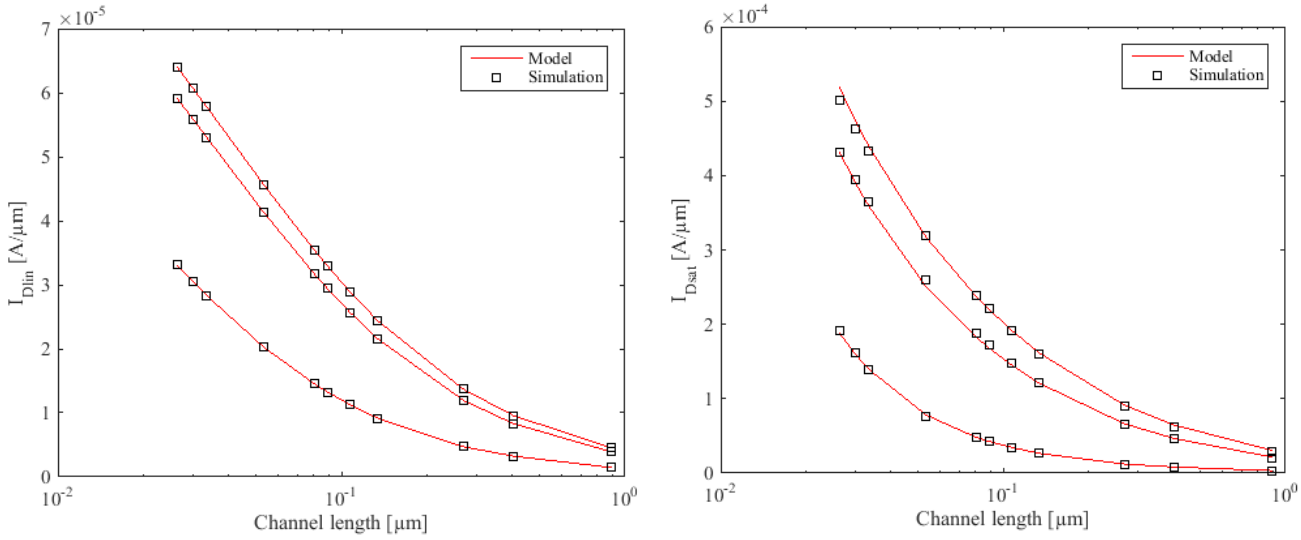
Tableau 2: Valeurs des paramètres utilisés pour le plan d'expérience en fonction du niveau considéré

Nous avons montré que la réponse des paramètres modèle aux variations du procédé est physiquement cohérente, attestant du sens physique des paramètres modèle et de la robustesse de l'extraction. Les extractions ont été exécutées pour des dispositifs nMOS et pMOS permettant une quantification de l'impact de la dose de dopant actif dans la région source-drain, ainsi que le profil de jonction, sur le courant de drain et les paramètres modèle.



(a) (b)

Figure 13: (a) R_0 extrait sur le plan d'expérience simulé en TCAD sur le pMOS. (b) R_0 extrait en fonction de la résistance R_{ext} ajouté aux contacts.

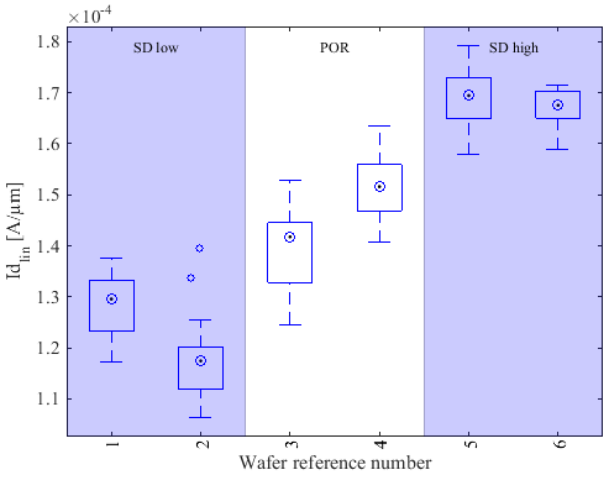


(a) (b)

Figure 14: Courant de drain linéaire (a) et saturé (b) simulé et modélisé pour le dispositif pMOS en fonction de L et V_G. Les différentes longueurs et tension V_G sont celles utilisées pour l'extraction.

Après l'introduction de la procédure d'extraction des paramètres modèle et son application sur des simulations TCAD dans le chapitre 3, nous l'avons appliqué sur des mesures sur silicium dans le chapitre 4, où les technologies 28 et 14 nm FD-SOI ont été étudiées. Il a été montré que les variations de paramètres modèles en fonction des variations du procédé sont cohérentes et ont été interprétées physiquement. Une quantification précise de l'impact des variations de du processus a été possible, ce qui montre que la méthode est efficace et robuste tout en ne nécessitant que peu de mesures, ce qui convient pour une application industrielle.

L'étude du 28 nm FD-SOI via l'extraction de paramètres du modèle a permis de quantifier l'impact de la dose implantée dans la région source-drain et de l'énergie d'implantation (voir Figure 4-3) ainsi que l'impact du DSA.



(c)

Figure 15: Distribution des courants de drain linéaire à $V_G = V_{dd}$ pour chaque wafer. Ici la dose et l'énergie d'implantation change

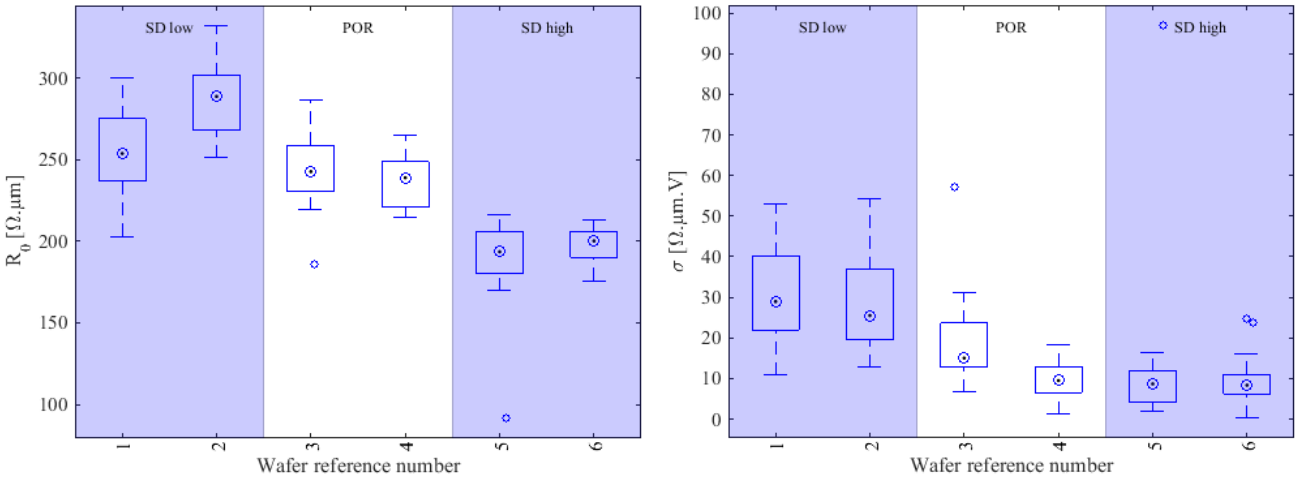


Figure 16: Distribution de R_0 (a) et σ (b) pour chaque wafer. Ici la dose et l'énergie d'implantation change.

Nous avons vu que les extractions donnent des résultats physiquement cohérents. La résistance de la région source-drain fortement dopée R_0 est abaissée par l'augmentation de la dose et de l'énergie de l'implant (voir Figure 4-4) ainsi que par l'augmentation de température du DSA. Ces deux paramètres du procédé influent directement sur la concentration de dopant actif. Cela signifie qu'il reste des dopants inactivés dans les régions source-drain avant le DSA. Le DSA les active avec succès. Au contraire, la résistance de la région faiblement dopée (sous l'espaceur), représentée par σ dans la formule de courant drain linéaire, ne dépend que de la dose de l'implant et de l'énergie (voir Figure 4-4). En effet, le DSA ne fait pas migrer les dopants et donc ne déplace pas la position de la jonction. De plus, cela signifie que les dopants dans la région faiblement dopée sont déjà bien activés avant le DSA et le DSA n'a pas d'effet d'activation dans cette région. Toutefois, l'extraction de Vt_{LDR} a mis en évidence que la position de la jonction est sensible à l'énergie d'implantation et à la dose. $\mu_0.C_{ox}$, θ_2 et Vt_{lin} se sont révélés être constant, ce qui signifie que le dopant ne pénètre pas dans la grille ou dans le canal. Toutes ces sensibilités peuvent être quantifiées facilement en utilisant cette technique, apportant des informations précieuses en termes d'optimisation de l'appareil.

L'étude de la technologie 14 nm FD-SOI a permis d'évaluer l'impact du temps de nettoyage HF avant épitaxie, la dose de carbone et de phosphore lors de l'épitaxie des régions source-drain dopées in situ ainsi que l'épaisseur d'épitaxie. Le carbone a induit une augmentation de R_0 en limitant la migration des dopants alors que l'accroissement de la dose de phosphore diminue R_0 en augmentant la dose de dopant actif dans la région source-drain fortement dopée. Un nettoyage court à l'HF avant l'épitaxie induit plus défauts qui agissent comme un puits de dopant, les empêchant de migrer vers le canal. La jonction est de fait éloignée du canal et la résistance de la région faiblement dopée augmente.

Dans une deuxième étape, la variabilité intra-wafer a été étudiée sur la technologie 14 nm FD-SOI. La méthode Monte Carlo ainsi que la propagation de variance avant (FPV) et arrière (BPV) ont été menées afin de modéliser cette variabilité. Les résultats sont montrés dans la Figure 4-19.

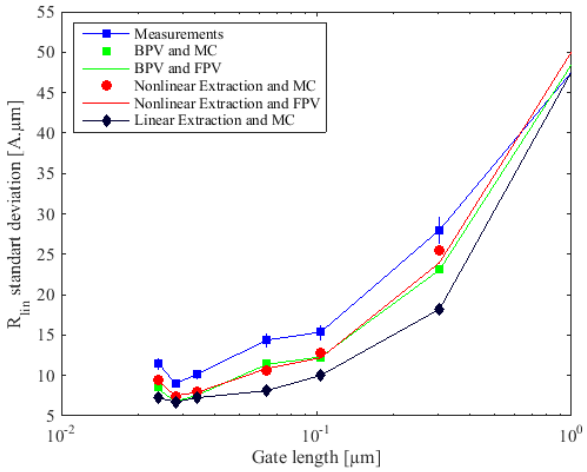
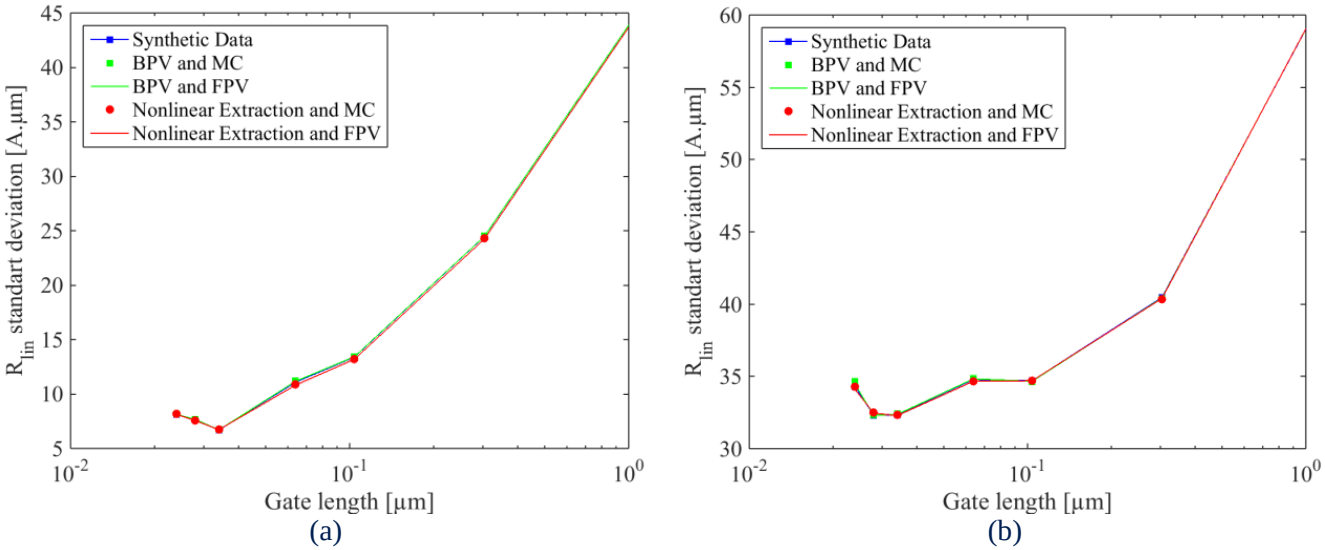
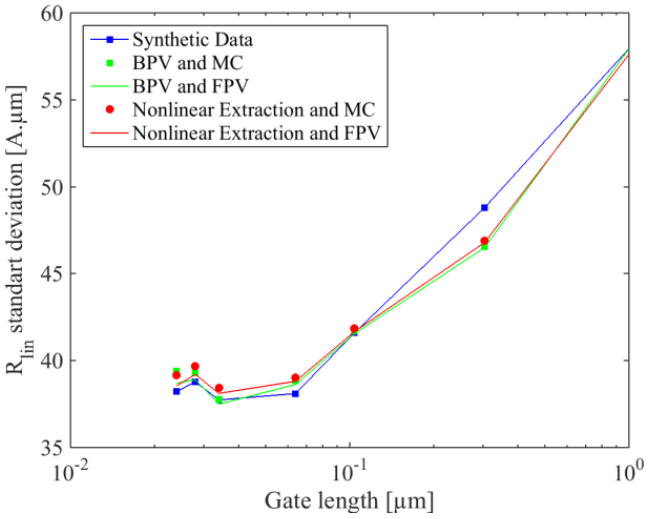


Figure 17: Ecart type de la résistance linéaire R_{lin} du nMOS sur la plaque de référence en fonction de la longueur de grille. Les résultats incluent l'étude Monté-Carlo, la FPV basée sur la dispersion des paramètres extraits par la méthode des moindres carrés linéaire et non linéaire ainsi que la BPV.

Il a été démontré que la variation de courant de drain linéaire est légèrement sous-estimée. La méthode BPV et l'extraction directe ont montré des résultats proches en termes de variabilité de courant de drain linéaire cependant la variabilité des paramètres modèle correspondants sont différents. Il a ainsi été suggéré que la variabilité locale et de la longueur du canal sont responsables de ces écarts (qui ne sont pas correctement pris en compte par extraction directe ou par BPV). Cette interprétation a été renforcée par le fait que la différence ne vient pas d'une violation des hypothèses normalité des distributions et de linéarité locale de la fonction. En effet, la méthode Monte Carlo a été utilisée pour propager la variabilité des paramètres modèle extrait en utilisant la méthode BPV et l'extraction directe, conduisant aux même résultats. Afin de vérifier que la longueur du canal et la variabilité locale sont responsables des écarts observés entre les mesures et le modèle, leur impact sur le modèle a été évalué à l'aide des données synthétiques, montrant qu'il induit des erreurs et peut donc expliquer ces écarts. Ces résultats sont montrés dans la Figure 4-22.





(c)

Figure 18: Variabilité des données synthétiques de résistance linéaire. Les données sont générées avec (a) ni la variabilité locale ni la variabilité de la longueur de grille, (b) la variabilité intra plaque de la longueur de grille et (c) la variabilité locale et la variabilité de la longueur de grille.

Le chapitre 5 présente la procédure pour construire et calibrer les polynômes qui lient les paramètres du procédé et modèle (la première étape de notre PCM, présentée sur la Figure 9). Dans le mesure où les paramètres du procédé sont nombreux et certains d'entre eux ne sont pas pertinents en fonction du paramètre modèle considéré, construire des modèles polynômiaux se confronte à deux problèmes: i) le problème est mal conditionnés et ii) il faut sélectionner les variables pertinentes. Ces problèmes sont adressés en utilisant des méthodes statistique appropriées, comme la régression pas à pas, la méthode LASSO ou LARS. Il a été démontré, en utilisant des données de synthèse, que ces méthodes sont en mesure d'effectuer la sélection de variables dans le cas de problèmes mal conditionnés et observations bruitées. L'exemple de l'application de la méthode LASSO est montrée sur la Figure 5-7:

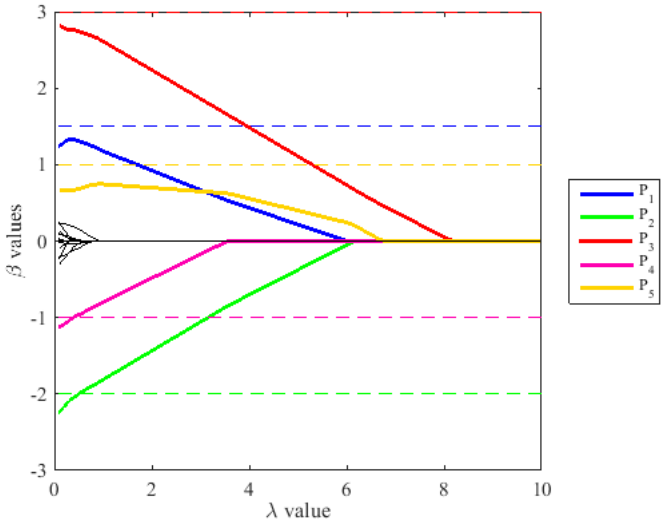
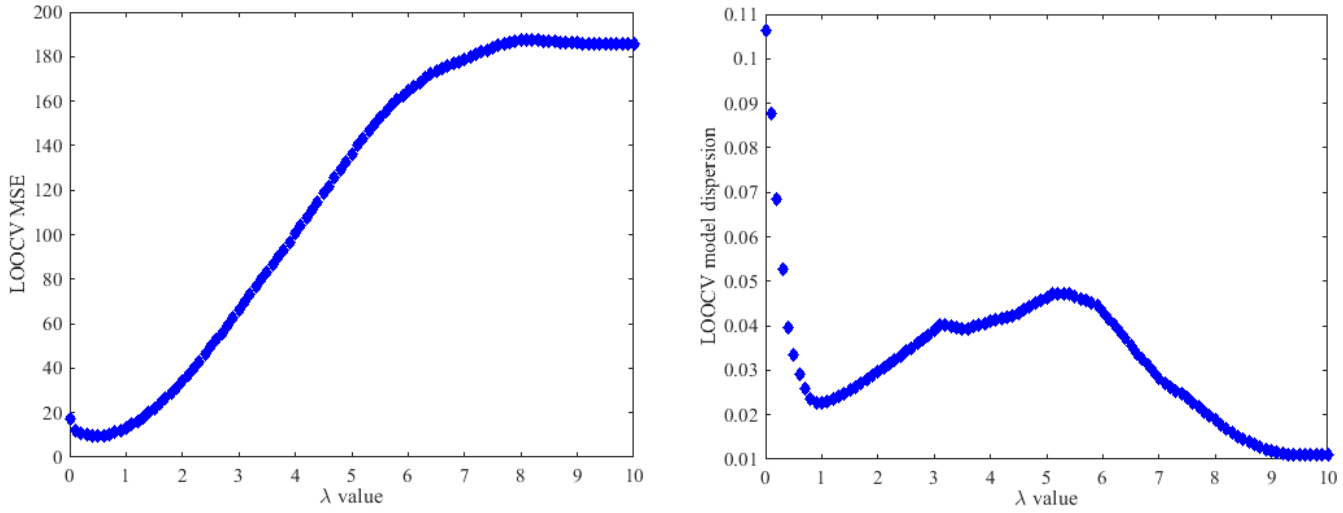


Figure 19: Valeurs de β extraites en fonction du paramètre λ choisie pour la méthode LASSO

L'application de la méthode LASSO donne les résultats montrés sur la Figure 5-9:

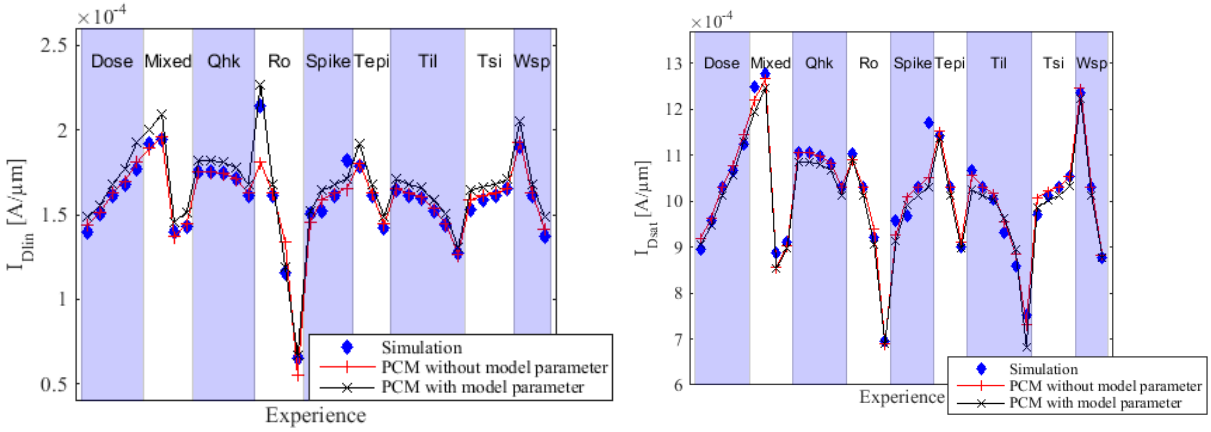


(a) (b)

Figure 20: Erreur moindre carré (a) et estimation de la variance (b) du modèle en fonction de la valeur du paramètre de calibration λ de la méthode LASSO.

L'application de la méthode LASSO suggère une valeur de calibration pour le paramètre λ de 1. Cette valeur est celle qui minimise à la fois l'erreur et la variance du modèle. Elle permet à la fois de supprimer les prédicateurs factices tout en gardant les prédicteurs pertinents avec des valeurs de coefficients proches de celles exactes.

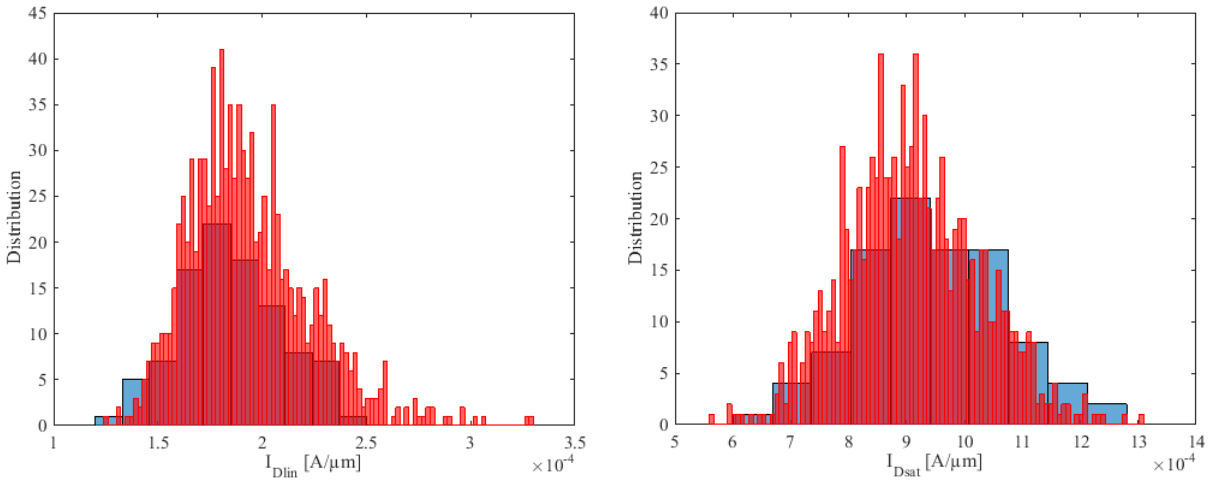
La procédure a été appliquée sur les résultats d'un plan d'expériences simulé en TCAD afin de tester sa fiabilité. Les modèles obtenus donnent des résultats précis et fiables comme le montre la Figure 5-21.



(a) (b)

Figure 21: Comparaison entre I_{Dlin} et I_{Dsat} modélisés par le PCM (avec et sans la phase modèle compact) et simulés en TCAD sur le plan d'expérience.

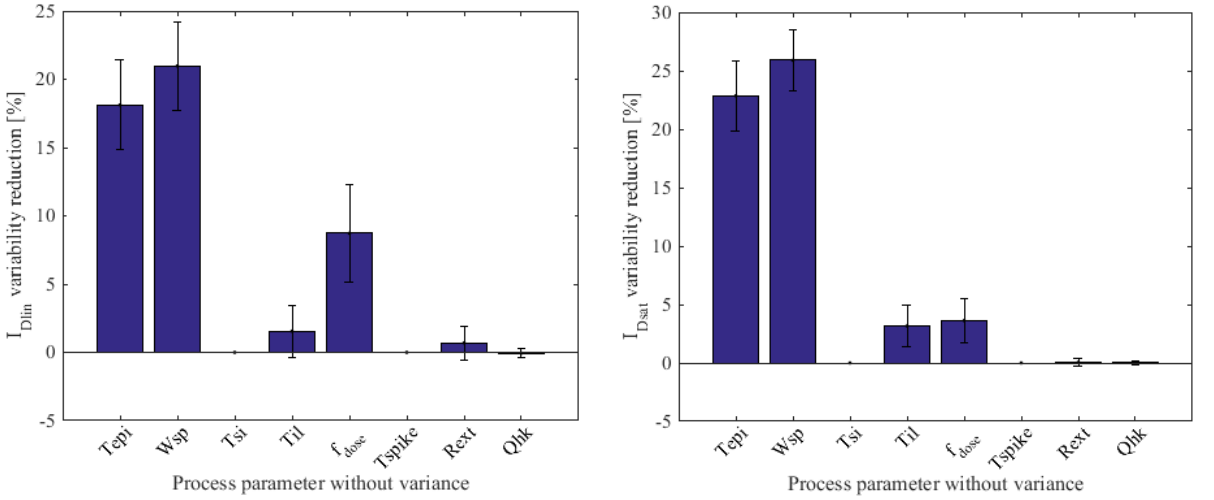
Ensuite, la variabilité intra-wafer simulée en TCAD a été modélisée en utilisant le PCM, montrant un bon accord comme le montre la figure Figure 5-32. Ces résultats montrent que le PCM est capable de modéliser la dispersion intra-plaque.



(a) (b)

Figure 22: Within-wafer distribution of I_{Dlin} and I_{Dsat} modeled and simulated using TCAD

Les paramètres du procédé ont ensuite été classés en fonction de leur contribution sur la variabilité du courant de drain (voir Figure 5-33).

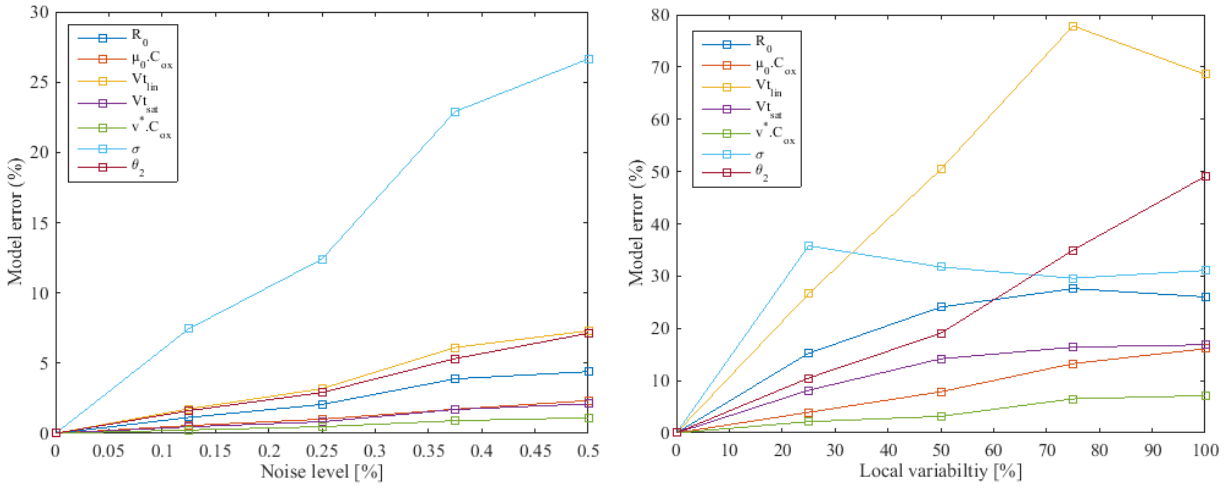


(a) (b)

Figure 23: Expected drain current variability reduction by suppressing the variability of one process parameter at a time.

Le modèle a montré que les paramètres Tepi et Wsp sont principalement responsables de la variabilité de I_{Dlin} et I_{Dsat}. Ainsi nous avons suggéré de réduire autant que possible la variabilité de ces paramètres afin de tirer le maximum de bénéfice en termes de variabilité de courant de drain.

Afin d'assurer la robustesse du processus de construction de PCM, les mesures électriques doivent répondre à des exigences spécifiques en termes de quantité et sur l'incertitude des mesures. Les limites de cette approche vis-à-vis de ces exigences ont été discutées. Nous avons montré que la construction du modèle est compromise par les effets du bruit et des variations locales, si un seul transistor est mesuré avec un court temps de mesure. Ces résultats sont montrés sur la Figure 24 :



Tableau

Figure 24: Erreur sur les paramètres modèles entre le PCM construit sur les données synthétiques et celui utilisé pour générer les données synthétiques en considérant (a) uniquement le bruit de mesure (b) uniquement la variabilité locale.

Cependant, nous avons montré que ce problème peut être surmonté en utilisant facilement une matrice de transistors. Nous recommandons d'utiliser des matrices de transistors 20x20, afin d'atteindre un niveau de bruit approprié. Cette mesure à elle seule résout aussi le problème du bruit dans ce cas d'étude. Il n'y a donc pas besoin d'augmenter la durée de la mesure. Bien entendu l'augmentation de la durée de la mesure renforcerait d'autant plus la robustesse de la construction de PCM. Les résultats sont montrés sur la Figure 25 :

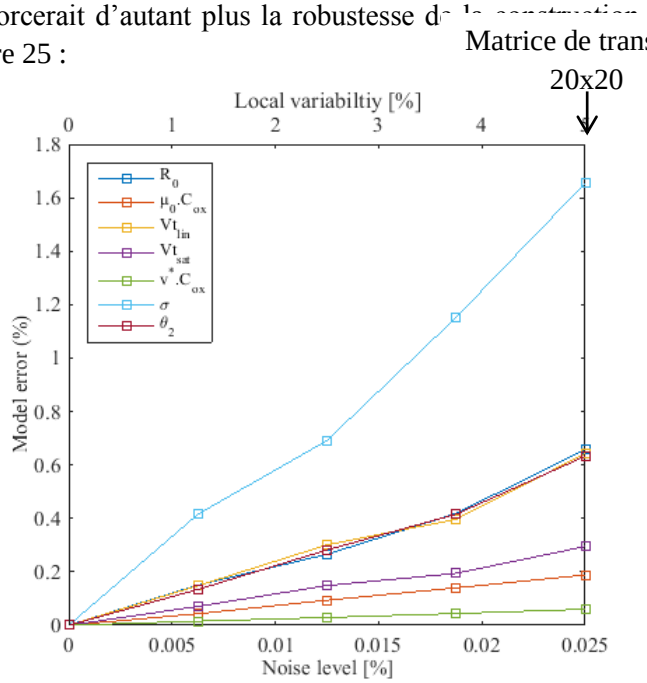


Figure 25: Erreur sur les paramètres modèles entre le PCM construit sur les données synthétiques et celui utilisé pour générer les données synthétiques en considérant le bruit de mesure et la variabilité locale

6.4 Applications et perspectives

Pour résumer, ce travail est une étude de faisabilité sur la construction de PCM et montre comment tirer parti de l'extraction à grande échelle des paramètres du modèle afin d'accélérer le du processus de développement. Nous avons montré que, avec très peu d'investissements, l'approche donne des résultats intéressants. En effet, seulement quelques points de mesure ont été utilisés à la place de caractérisation I_D-V_G complètes, traditionnellement utilisées pour la calibration du modèle et ceux sans aucune structure de test spécifique. Chaque algorithme a été exécuté à l'aide d'un ordinateur portable

avec une puissance moyenne de traitement, associé avec le logiciel flexible, mais plutôt lent, Matlab. Sur cette base, des conclusions sur l'effet du processus de fabrication sur les performances électriques ont été établis et le PCM a été construits sur la base d'un plan d'expériences simulé en TCAD. Le PCM a été en mesure de fournir des lignes directrices afin d'optimiser la variabilité du courant de drain.
La qualité et la quantité des bénéfices tirés sont proportionnelles à la quantité de ressources investies. En fait, il y a un compromis en graduel entre robustesse et flexibilité du modèle, et la quantité nécessaire de ressources à investir. Dans les sections suivantes, nous examinons les bénéfices potentiels qui pourraient être tirés en utilisant le PCM avec quelques fonctionnalités avancées.

6.4.1 Optimiser le du processus de fabrication

L'optimisation de la variabilité via le PCM a été étudiée sur le silicium, la TCAD et les données synthétiques. Cependant, il a été suggéré que cette procédure peut être appliquée pour optimiser la performance et la variabilité dans le même temps, à condition que le modèle soit bien calibré. Pour atteindre cet objectif, nous proposons ici, en guise d'application, une procédure générale visant à optimiser les performances et la variabilité en même temps. Cette procédure permet également de calibrer l'outil de simulation TCAD. La description de la procédure est représentée sur la Figure 26.

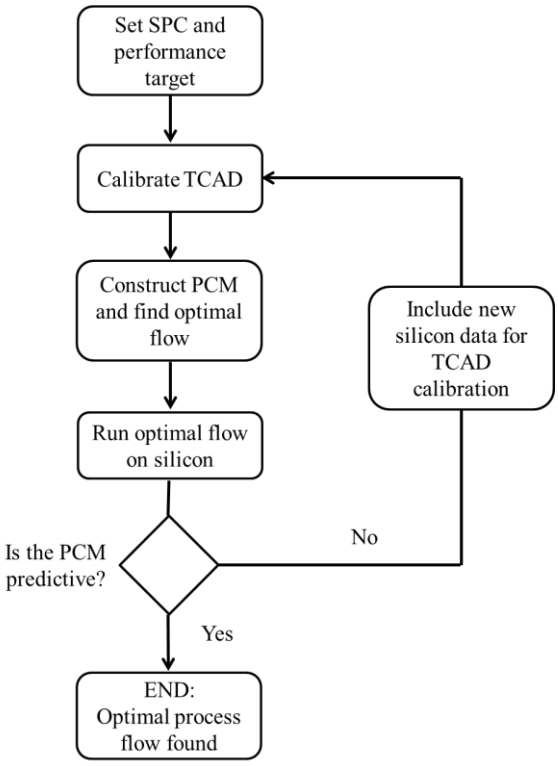


Figure 26: Organigramme de la procédure d'optimisation de la performance et de la variabilité

La Figure 26 montre l'organigramme de la procédure d'optimisation de la performance et de la variabilité. Dans un premier temps, cette procédure consiste à définir des objectifs en termes de performance et de variabilité sur les paramètres électriques. Ensuite, les paramètres du simulateur TCAD doivent être calibrés. Cette procédure peut être faite en utilisant le modèle compact ainsi que sa procédure d'extraction. Elle est détaillée dans la Figure 27.

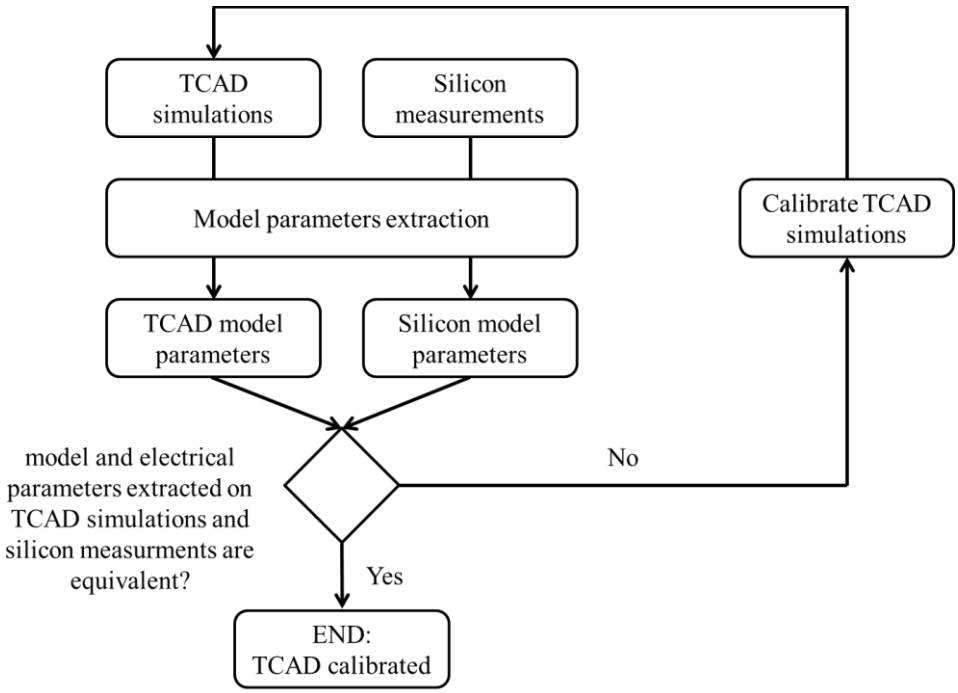


Figure 27: Organigramme de la procédure de calibration de l'outil TCAD

Dans ce mode opératoire, on extrait les paramètres du modèle à l'aide de la TCAD et de mesures sur silicium. La comparaison des paramètres électriques et modèles obtenus sur la TCAD et le silicium permet d'évaluer la précision de la calibration TCAD. Si la TCAD n'est pas correctement calibrée, le décalage entre les paramètres du modèle de la TCAD et du silicium donne des indications pour re-calibrer la TCAD correctement. Par exemple, s'il y a une bonne adéquation entre tous les paramètres du modèle sur la TCAD et le silicium à l'exception de $\mu_0.C_{ox}$, alors le modèle de mobilité de la TCAD et/ou l'épaisseur d'oxyde équivalente du dispositif devraient être étudiées pour la calibration de la TCAD.

Lorsque la TCAD est calibré, le PCM doit être construit (suivant l'organigramme de la Figure 26). La procédure de construction du PCM est détaillée sur l'organigramme de la Figure 28. La procédure consiste à simuler un plan d'expériences, extraire les paramètres du modèle à partir des courants de drain simulés et construire le PCM suivant les instructions détaillées au chapitre 5. Le procédé de fabrication optimal est alors trouvé grâce à un algorithme d'optimisation non linéaire se basant sur le PCM. La pertinence des résultats doit être vérifiée après coup. En effet, si l'on considère le PCM construire au chapitre 5, nous pouvons voir que R_0 est linéairement proportionnel à la dose de l'implant. Ainsi, si nous tenons à optimiser le procédé de fabrication de tel sort qu'il maximise le courant de drain, alors la solution suggèrera d'augmenter la dose implantée indéfiniment de telle sorte que R_0 est minimisée. En pratique, on sait qu'il y a une concentration de dopant maximale au-dessus de laquelle aucun gain en résistance d'accès n'est attendu, car il y a un effet de saturation de la concentration en dopant. En d'autres termes, le domaine de validité du PCM est trop étroit et non convexe. De fait la solution optimale trouvée peut ne pas être bornée.

Pour corriger ce défaut, le dispositif doit être étudié lorsque les paramètres du procédé atteignent des valeurs extrêmes. Pour cet exemple, les simulations doivent être exécutées avec des doses d'implants suffisamment élevées afin de capturer l'effet de saturation de la concentration de dopants actif. Lorsque le nouveau plan d'expériences est conçu, la procédure de simulations et de construction du PCM doit être exécutée avant d'optimiser le procédé de fabrication une nouvelle fois. Cette boucle doit être répétée jusqu'à ce qu'une solution physiquement pertinente et bornée soit trouvée pour le procédé de fabrication.

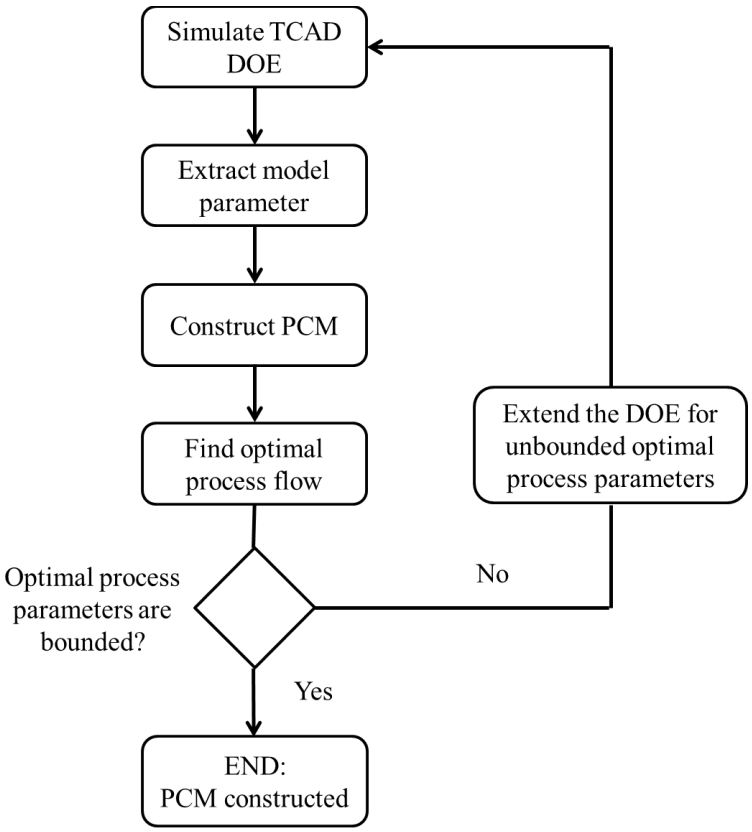


Figure 28: Organigramme de construction et de calibration du PCM

Suivant l'organigramme de la Figure 26, une dernière vérification doit être effectuée de manière à déterminer si le du processus de fabrication trouvé par l'optimisation est en fait la solution optimale. Ce test consiste à exécuter le du processus de fabrication optimal sur silicium. Les résultats trouvés en utilisant le PCM et les mesures silicium doivent ensuite être comparées. Si le modèle ne correspond pas à silicium, cela signifie que, soit la TCAD est mal étalonnée (compte tenu de ce nouveau du processus de fabrication) ou le PCM n'est pas suffisamment prédictif. Toute la procédure doit être exécutée une fois de plus en se concentrant maintenant sur ce nouveau du processus de fabrication pour l'étalonnage du modèle et de la TCAD.

Toute la procédure, comme le montre la Figure 26, est itérative et quelques itérations pourraient être nécessaires afin d'arriver à un modèle cohérent et un du processus de fabrication optimal. Il n'est nécessaire de traiter que quelques plaques par itération, ce qui rend l'approche très rentable. En outre, une procédure entièrement automatisée peut effectuer une itération très rapidement (de l'ordre de grandeur de quelques heures). La seule étape qui ne peut pas être automatisé est calibration TCAD car elle nécessite l'expertise d'ingénieurs qualifiés. Cependant les indications fournies par l'extraction des paramètres du modèle peuvent grandement faciliter cette tâche.

6.4.2 Fonctionnalités avancées pour les futures études basée sur l'outil PCM

En perspective, nous proposons ici quelques lignes directrices afin d'améliorer l'approche développée dans ce travail et de tirer pleinement profit de la technique. La Figure 29 représente le schéma du PCM comme celui présenté dans l'introduction. Cependant ici, nous avons ajouté des fonctionnalités avancées qui pourraient être étudiées dans les applications futures.

Les fonctionnalités avancées incluent l'utilisation de nouveaux types de modèle pour relier les paramètres du procédé aux paramètres modèle. Parmi ces nouveaux types on trouve, les réseaux de neurones (Feed Forward Neural Network) **[156][186][187]**, le Support Vector Machine (SVM) **[188]** ou simplement un modèle physique (possiblement non-linéaire) défini par l'utilisateur. Ces nouvelles

approchent contrecarrent les limitations de l'approche actuelle, basée sur des polynômes linéaires, qui n'incluent pas de deuxième ordre et ni les effets croisés des paramètres du procédé.

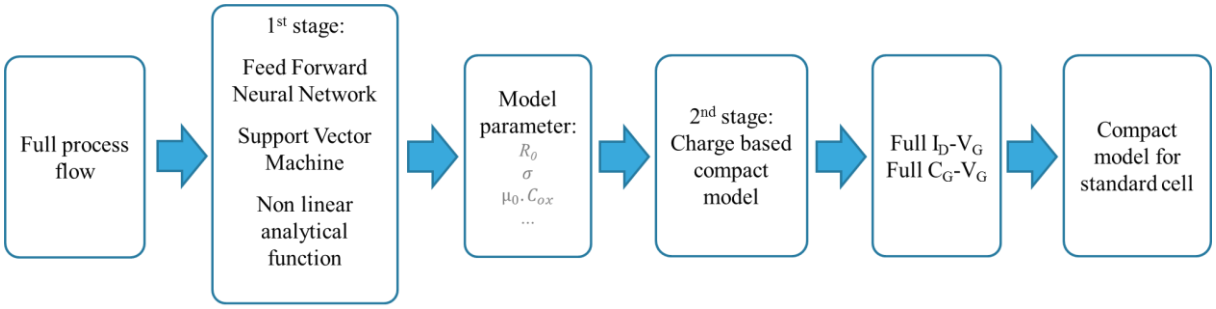


Figure 29: Organigramme du PCM pour cellules standard

En outre, la dépendance non linéaire entre certains paramètres comme sigma vs Tspike, ou difficile à modéliser avec des polynômes linéaires. Enfin, afin de trouver le procédé de fabrication optimale, nous avons vu que le PCM doit être valable sur de larges variations de paramètres du procédé et doit donc prendre en compte les relations non linéaires entre le modèle et les paramètres du procédé (tels que la saturation de la concentration de dopants pour les doses d'implants très élevés). Cela ne peut pas être pris en compte par l'utilisation de polynômes linéaires mais des formules non linéaires peuvent prendre en compte ce genre de comportement.

Les méthodes FFNN et SVM peuvent également prendre en compte ces non linéarités. En outre, ces méthodes peuvent travailler avec des paramètres du procédé discontinus (par exemple des variables booléennes). En conséquence, il peut gérer des changements dans le procédé de fabrication (suppression ou addition d'étapes, changement d'outil). Cependant, considérant la méthode FFNN, il convient de noter que cette méthode est moins transparente par rapport aux polynômes, bien qu'elle soit très facile à manipuler et à former. De plus, c'est une méthode très puissante capable de modéliser des systèmes complexe, de nature très variée. Voilà pourquoi elle est souvent appelée «méthode d'approximation universel». Aujourd'hui, le réseau de neurones artificiel trouve un nombre croissant d'applications, allant de la reconnaissance faciale ou de la parole, au diagnostic médical, pour n'en nommer que quelques-unes.

Les fonctionnalités avancées citées précédemment incluent également l'utilisation d'un modèle compact plus souple et précis. Ceci est suggéré sur la Figure 29 en utilisant un modèle compact basé sur la charge ou sur le potentiel de surface. La principale ligne directrice que je fournirais est d'utiliser un modèle compact avec des paramètres qui ont une signification physique claire et la plus élémentaire possible. Cet atout simplifierait grandement la première étape du PCM et rend l'ensemble du PCM beaucoup plus robuste. En outre, nous pourrions espérer modéliser les caractéristiques I_D-V_G complètes, voir même les caractéristique C_G-V_G. Bien entendu, la limitation principale est d'avoir un nombre limité de paramètres modèle (environ 10), afin de pouvoir les extraire avec peu de mesures. Enfin, la dernière option à étudier est d'étendre le modèle en ajoutant une troisième étape. Cette étape modéliserait les caractéristiques électriques de cellules standards à l'aide de celles du transistor isolé (par exemple la SNM pour la SRAM ou la vitesse de commutation pour un oscillateur en anneau). Cette troisième étape peut être extrêmement précieuse. En effet, dans cette thèse, nous avons mis l'accent sur l'optimisation des performances et de la variabilité des paramètres électriques I_{Dlin} et I_{Dsat}. Les cibles à atteindre (en termes de courant de drain) sont définies de sorte qu'il assure les fonctionnalités de circuit. Cependant, avec un PCM capable de modéliser les performances des cellules standards, il serait possible d'optimiser directement les performances de la cellule standard. Cette approche offrirait une plus grande liberté et une étendue plus large de solutions en termes de procédé de fabrication optimal. Dans une autre mesure, il serait également possible de considérer l'effet du tracé du circuit avec ce type PCM. Ainsi l'optimisation ne serait pas limitée à trouver le

procédé de fabrication optimal, mais aussi le tracée du circuit optimal. Ce type de procédure d'optimisation globale donnerait des solutions à forte valeur ajoutée.

Figure 30 expose un du processus pour la construction de PCM avec des fonctionnalités avancées qui méritent d'être étudiées dans les travaux futures.

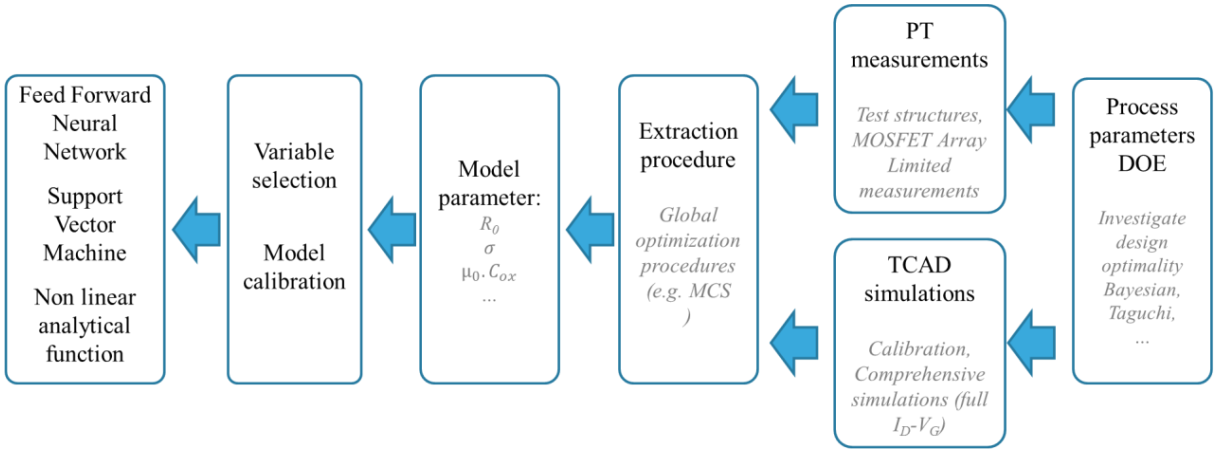


Figure 30: Organigramme simplifié de la construction du PCM avec les fonctionnalités avancées

Dans cette figure, le flux de construction commence par la construction d'un plan d'expériences approprié visant à étudier les effets des paramètres du procédé. Selon le modèle que nous essayons de construire (en particulier pour la première étape PCM), différents type de plan d'expériences peuvent être utilisés. Les méthodes de construction de plan d'expériences optimal ont été largement étudiées. Les critères universels (basés sur la matrice d'information de Fisher) pour construire des plans d'expériences optimaux ont été introduites par Wald (1943), Elving (1952), Kiefer (1959) et Kiefer (1975). Une littérature vaste et complète peut être trouvée sur ce sujet **[311]**. Choisir le bon plan d'expériences minimise la chance de construire des PCMs inexactes et augmente sa robustesse. Par exemple, afin de construire un modèle polynômial du second d'ordre, un plan d'expériences composite centré serait préférable au plan d'expériences qui a été utilisé pour les modèles polynômiaux linéaire.

Une autre caractéristique qui permettrait d'améliorer la technique consiste à utiliser des méthodes d'optimisation globale afin d'améliorer la procédure d'extraction. Dans notre approche, nous avons utilisé la méthode dite « trust-reflective-region » avec l'algorithme du gradient conjugué. Il est très efficace compte tenu de notre problème, mais il ne peut pas assurer de trouver l'optimum global du problème, surtout si la première hypothèse est mauvaise. Dans ce travail, nous avons contourné ce problème en utilisant la première estimation trouvée avec un ajustement par moindres carrés linéaire. Cette approche garantit d'avoir une première estimation assez proche de l'optimum global. En outre, la robustesse de l'extraction a été largement testée et éprouvé. Donc, compte tenu de notre cas, la méthode d'optimisation fonctionne très bien. Mais si un modèle compact plus précis est utilisé pour lesquels aucune première estimation ne peut être fournie avant l'exécuter de l'étape d'optimisation, alors un algorithme d'optimisation globale pourrait être avantageux. Une large gamme de solutions existe et certains d'entre eux ont été étudiés dans la littérature pour l'extraction de paramètres **[133][134][314][315]**. Différents algorithmes ont été testés au cours de cette thèse (comme algorithme génétique et Levenberg-Marquardt), et nous vous recommandons d'utiliser les méthodes basées sur la dérivée de la fonction objective, car les modèles compacts sont continus, dérivables et rapide à calculer. Ces méthodes semblent être plus rapides et plus précises. Dans cette perspective, nous pouvons mentionner la méthode « Multi-levels coordinates search » (MCS) comme alternative à notre approche **[317]**. Plus d'informations sur l'optimisation globale peut être trouvée dans la littérature **[318]**.

Si l'on devait utiliser les méthodes FFNN, SVM ou des formules analytiques non linéaires pour la première étape du PCM, alors des méthodes spécifiques doivent être utilisées pour la sélection de variables et l'étalonnage du modèle. Les méthodes FFNN et SVM n'exigent pas explicitement

l’utilisation des méthodes de sélection des variables, mais elles peuvent améliorer leur efficacité pour la calibration et l’utilisation des modèles. La littérature rapports de multiples méthodes pour effectuer cette sélection de variables conformément à ces méthodes **[319] - [323]**. Dans le cas où l'on veut utiliser des modèles non linéaire, d'autres méthodes sont plus appropriées pour effectuer la sélection de variables **[180] - [330]**.

6.4.3 Fonctionnalités avancées pour les futures études basée sur l’outil PCM

En termes d'applications inexplorées nous avons déjà mentionné la possibilité d'optimiser les performances des cellules standards au lieu des transistors. Une autre application peut être suggérée. Jusqu'à présent, nous avons décrit une technique pour trouver le procédé de fabrication optimal. Cependant, au cours du processus, la performance et la variabilité peut être impacté par la dérive de l'étalonnage des outils. En d’autres termes, l'étalonnage des outils peut dériver avec le temps et modifier légèrement la valeur moyenne des paramètres du procédé sur la plaque. En fin de compte, il y aura des divergences entre le procédé de fabrication optimal et celui qui a été effectivement effectué. Le même problème peut être observé à l'échelle de la puce. En effet, les paramètres du procédé ne sont pas répartis uniformément sur la plaque, en raison de la variabilité au sein de celle-ci. Souvent, il existe une signature de plaque (par exemple, un gradient radiale ou linéaire) de la dispersion de ces paramètres du procédé. Ainsi, le du processus observé à l'échelle de la puce peut être différent du processus de fabrication optimal. Pour contrecarré ce problème, une étude a été faite sur l’ajustement du processus in situ, afin de réduire l'effet de la dérive du procédé **[307]**. La méthode consiste à faire, en temps réel et en ligne, la surveillance des du processus afin d'estimer leur dérive à l'échelle de la puce. Puis à une étape critique, le du processus de fabrication peut être réajusté grâce à un recalibrage de l'outil ou une réorientation de la plaque afin de contrebalancer l'effet de la dérive de l'étalonnage de l'outil ainsi que de signature de la dispersion des paramètres du procédé sur la plaque. Cette requalification du processus peut être faite en utilisant le PCM. En effet, à l'étape critique du processus, au lieu d'optimiser le procédé de fabrication entier comme nous le suggère l'application précédente, seules les étapes restantes seraient optimisées, connaissant l'historique du processus. Ce traitement in situ et en temps réel de l'optimisation du processus grâce au PCM permettrait la maximisation du rendement.

Références du résumé en Français

[209] M. J. Sherony, L. T. Su, J. E. Chung, D. A. Antoniadis, “SOI MOSFET effective channel mobility”, IEEE Trans. Elec. Dev., Vol. 41, no. 2, pp. 276-278, 1994

[210] S. Takagi, A. Toriumi, M. Iwase, H. Tango, “On the universality of inversion layer mobility in Si MOSFET’s: Part I-Effects of substrate impurity concentration”, IEEE Trans. Elec. Dev., Vol. 41, no. 12, pp. 2357, 1994.

[211] D. Rideau, Y. M. Niquet, O. Nier, A. Cros, J.P. Manceau, P. Palestri, D. Esseni, V. H. Nguyen, F. Triozon, J.C. Barbé, I. Duchemin, D. Garetto, L. Smith, L. Silvestri, F. Nallet, R. Clerc, O . Weber, F. Andrieu, E. Josse, C. Tavernier, H. Jaouen, “Mobility in High-K Metal Gate UTBB-FD-SOI Devices: from NEGF to TCAD perspectives”, Proc. on International Electron Device Meeting (IEDM), pp. 12.5.1-12.5.4, 2013.

[212] G. Ghibaudo, M. Mouis, L. Pham-Nguyen, K. Bennamane, I. Pappas, A. Cros, G. Bidal, D. Fleury, A. Claverie, G. Benasayag, P-F. Fazzini, C. Fenouillet-Beranger, S. Monfray, F. Boeuf, S. Cristoloveanu, T. Skotnichi, N. Collaert, “Electrical transport characterization of nano CMOS devices with ultra-thin silicon film”, Ext. Abs. 9th international workshop on junction technology, pp. 58-63, 2009.

[213] V. Barral, T. Poirroux, D. Munteanu, J-L. Autran, S. Deleonibus, “Experimental investigation on the quasi-ballistic transport: Part II-backscattering coefficient extraction and link with the mobility.”, IEEE Trans. Elec. Dev., Vol. 56, no.3, pp. 420-430, 2009.

[214] I. Pappas, G. Ghibaudo, C. A. Dimitriadis, C. Fenouillet-Beranger, “Backscattering coefficient and drift-diffusion mobility extraction in short channel MOS devices”, Solid States Electronics, Vol. 53, pp. 54-56, 2009.

[215] S. Guarnay, F. Triozon, S. Martinie, Y. M. Niquet, A. Bournel, “Monte Carlo study of effective mobility in short channel FD-SOI MOSFETs”, Proc. of SISPAD international conference, 2014

[216] D. Fleury, G. Bidal, A. Cros, F. Boeuf, T. Skotnicki, G. Ghibaudo, “New Experimental insight into ballistic of transport in strained bulk MOSFETs”, Proc. of Symposium on VLSI technology, pp. 16-17, 2009.

[217] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, S. Mukhopadhyay, “Magnetoresistance mobility characterization in advanced FD6SOI n-MOSFETs”, Solid States Electronics, Vol. 10, pp. 229-235, 2015.

[218] M. Zilli, P. Palestri, D. Esseni, L. Selmi, “On the experimental determination of channel backscattering in nano MOSFETs.”, IEDM Tech Digest, pp.105, 2007.

[219] K. Huet, J. Saint-Martin, A. Bournel, S. Galdin-Retailleau, P. Dollfus, G. Ghibaudo, M. Mouis, “Monte Carlo study of apparent mobility reduction in nano-MOSFETs”, Proc of ESSDERC, pp. 382-385, 2007

[220] E. J. Ryder, “Mobility of Holes and Electrons in High Electric Fields”, Physical Review, Vol. 90, no. 5, pp. 766-769, 1953.

[221] C. B. Norris, J. F.Gibbons, “Papers on Carrier Drift Velocities in Silicon at High Electric Field Strengths”, IEEE Trans. Elec. Dev., Vol. 14, no. 1, pp.37, 1967.

[222] C. Y. Duh, J. L. Moll, “Electron Drift Velocity in Avalanching Silicon Diodes”, IEEE Trans. Elec. Dev., Vol. 14, no. 1, pp.46, 1967.

[223] V. Rodriguez, H. Ruegg, M-A. Nicolet, “Measurement of the drift velocity of holes in silicon at high-field strengths”, IEEE Trans. Elec. Dev., Vol. 14, no. 1, pp.44, 1967.

[224] J. G. Ruch, “Electron Dynamics in Short Channel Field-Effect Transistors” , IEEE Trans. on Elec. Dev., Vol. 19, no. 5, pp. 652-654, 1972.

[225] J. Kim, J. Lee, Y. Yun, B-G. Park, J. D. Lee, H. Shin, “Extraction of Effective Carrier Velocity and Observation of Velocity Overshoot in Sub-40 nm MOSFETs”, Journal of semiconductor technology and science, Vol. 8, no.2, pp.115-120, 2008.

[226] M. Lundstrom, “Elementary Scattering Theory of the Si MOSFET”, IEEE Elec. Dev. Lett., Vol. 18, no. 7, pp. 361-363, 1997

[227] Peizhen Yang, W.S. Lau, Seow Wei Lai, V.L. Lo, S.Y. Siah and L. Chan (2010). The Evolution of Theory on Drain Current Saturation Mechanism of MOSFETs from the Early Days to the Present Day, Solid State Circuits Technologies, Jacobus W. Swart (Ed.), ISBN: 978-953-307-045-2, InTech, DOI: 10.5772/6873. Available from: http://www.intechopen.com/books/solid-state-circuits-technologies/the-evolution-of-theory-on-drain-current-saturation-mechanism-of-mosfets-from-the-early-days-to-the-

[228] K. Natori, “Ballistic metal-oxide semiconductor field effect transistor”, Journal of Applied Physics, Vol. 76, no. 8, pp. 4879-4890, 1994.

[229] C. C. Hu, “MOS Transistor”, in Modern Semiconductor Devices for Integrated Circuits, 1st Ed., Prentice Hall, 2010, ch. 6, sec. 6.3.1, pp. 202.

[230] G. Ghibaudo, “Analytical modeling of the MOS transistor”, Phys. Stat. Sol., Vol. 113, pp. 223-239, 1989.

[231] G. Ghibaudo, “A simple model of the drain saturation voltage dependence with gate voltage for short channel MOSFETs”, Phys. Stat. Sol., Vol. 99, pp. K149-K153, 1987.

[232] L. Pham-Nguyen, C. Fenouillet-Beranger, A. Vandooren, A. Wild, G. Ghibaudo, S. Cristoloveanu, “Direct comparison of Si/High-K and Si/SiO2 channels in advances FD SOI MOSFETs”, Proc. of IEEE International SOI conference, pp. 25-26, 2008.

[233] M. Cassé, F. Rochette, N. Bhouri, F. Andrieu, D. K. Maude, M. Mouis, G. Reimbold, F. Boulanger, “Mobility of strained and unstrained short channel FD-SOI MOSFETs: new insight by magnetoresistance”, Proc. of Symposium on VLSI technology digest of technical papers, pp. 170-171, 2008.

[234] W. Chaisantikulwat, M. Mouis, G. Ghibaudo, C. Gallon, C. Fenouillet-Beranger, D.K. Maude, T. Skotnicki, S. Cristoloveanu, ”Magnetoresistance technique for mobility extraction in short channel FD-SOI transistors”, Proc. of IEEE ESSDERC, pp. 569-572, 2005.

[235] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G. Kim, G. Ghibaudo, “In depth characterization of electron transport in 12nm FD-SOI CMOS”, Solid States Electronics (2015), http://dx.doi.org/10.1016/j.sse.2015.02.012.

[236] S. Morvan, F. Andrieu, M. Cassé, O. Weber, N. Xu, P. Perreau, J. M. Hartmann, J. C. Barbé, J. Mazurier, P. Nguyen, C. Fenouillet-Beranger, C. Tabone, L. Tosti, L. Brévard, A. Toffoli, F. Allain, D. Lafond, B. Y. Nguyen, G. Ghibaudo, F. Boeuf, O. Faynot, T. Poirroux, “Efficiency of mechanical stressors in planar FD-SOI n and p MOSFETs down to 14nm gate length”, Proc. of Symposium on VLSI technology digest of technical papers, pp. 111-112, 2012.

[237] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G.-T. Kim, “Low temperature characterization of mobility in advanced FD-SOI n-MOSFETs under interface coupling conditions”, Proc on ULtimate Iintegrated on Silicon conference, pp. 61-64, 2014.

[238] S. R. Hofstein, F. P. Heiman, “Insulated-gate field effect transistor”, Proceedings of the IEEE, Vol. 51, no. 9, pp. 1190-1202, 1963.

[239] G. Merckel, J. Borel, N. Z. Cupcea, “An Accurate Large-Signal MOS Transistor Model for se in Computer-Aided Design”, IEEE Trans. Elec. Dev., Vol. ED19, no. 5, pp. 681-690, 1972.

[240] P. I. Suciu, R. L. Johnston, “Experimental Derivation of the Source and Drain Resistance of MOS Transistors”, IEEE Trans. Elec. Dev. Vol. ED27, no. 9, pp. 1846-1848, 1980.

[241] B. Cabon-Till, G. Ghibaudo, S. Cristoloveanu, “Influence of source drain series resistance on MOSFET Field-effect mobility”, Electronics Letters, Vol. 21, no. 11, pp. 457-458, 1985.

[242] G. J. Hu, C. Chang, Y. T. Chia, “Gate-Voltage-Dependent Effective Channel Length and Series Resistance of LDD MOSFET’s”, IEEE Trans. Elec. Dev. Vol. ED34, no. 12, pp. 2469-2475, 1987.

[243] K. K. Ng, W. T. Lynch, “Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFET’s”, IEEE Trans. Elec. Dev. Vol. ED33, no. 7, pp. 965-972, 1986.

[244] V. G. K. Reddi, C. T. Sah, “Source to Drain Resistance Beyond Pinch-Off in Metal-Oxide-Semiconductor Transistors (MOST)”, IEEE Trans. Elec. Dev. Vol. 12, no. 3, pp. 139-141, 1965.

[245] F. Monsieur, Y. Denis, D. Rideau, J. Lacord, V. Quenette, G. Gouget, C. Tavernier, H. Jaouen, ‘The importance of the spacer region to explain short channels mobility collapse in 28 nm Bulk and FD-SOI technologies’, IEEE Proc. on ESSDERC 2014.

[246] K. Y. Lim and X. Zhou, “A Physically-Based Semi-Empirical Series Resistance Model for Deep-Submicron MOSFET I–V Modeling”, IEEE Trans. Elec. Dev., Vol. 47, no. 6, pp. 1300-13012, 2000.

[247] B. J. Sheu, C. Hu, P. K. KO and F. C. Hsu, “Source-and-Drain Series Resistance of LDD MOSFET’s”, IEEE Elec. Dev. Lett., Vol. EDL-5, no. 49, pp. 365-367, 1984.

[248] S. D. Kim, C. M. Park, J. C. S. Woo, “Advanced Model and Analysis of Series Resistance for CMOS Scaling Into Nanometer Regime—Part I: Theoretical Derivation”, IEEE Trans. Elec. Dev. Vol. 49, no. 3, pp. 457-466, 2002.

[249] Y. Taur, “MOSFET channel length: extraction and interpretation”, IEEE Trans. Elec. Dev., Vol. 47, no. 1, pp. 160-170, 2000

[250] J. Kim, J. Lee, I. Song, Y. Yun, J. D. Lee, B-G. Park, H. Shin, “Accurate extraction of effective channel length and source/drain resistance on ultrashort channel MOSFETs by iteration method”, IEEE Trans. Elec. Dev., Vol. 55, no. 10, 2008.

[251] M.F. Hamer, “First-order parameter extraction on enhancement silicon MOS transistors”, IEEE Proc., Vol. 133, Pt. 1, no. 2, pp. 49-54, 1986.

[252] Q. Chen, E. M. Harrell, and J. D. Meindl, “A physical short-channel threshold voltage model for undoped symmetric double-gate MOSFETs,” IEEE Trans. Elec. Dev., vol. 50, no. 7, pp. 1631–1637, Jul. 2003.

[253] Y. P. Tsividis, C. McAndrew, “Operation and Modeling of the MOS Transistor”, Mc Graw-Hill Book Company, New York, p.155, (1987).

[254] H.S.P. Wong, M.H. White, T.J. Krutsick, R.V. Booth “Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFETs” , Solide-State Electronics, Vol. 30, no. 10, pp 953-968, 1987.

[255] J. Lacord, J. L. Huguenin, T. Skotnicki, G. Ghibaudo, F. Boeuf, “and Efficient MASTAR Threshold Voltage and Subthreshold Slope Models for Low-Doped Double-Gate MOSFET”, IEEE Trans. Elec. Dev., Vol. 59, no. 9, 2012.

[256] F. Balestra, I. Hafez, G. Ghibaudo, “A new method for the extraction of MOSFET parameters at ambient and liquid helium temperatures”, Journal de physique, Colloque C4, supplement au no. 9,

Tome 49, pp 817-820, 1988.

[257] A. Ortiz-Conde, F.J. Garcia Sanchez, J.J. Liou, A. Cerdeira, M. Estrada, Y. Yue, "A review of recent MOSFET threshold voltage extraction methods", Microelectronics reliability, Vol. 42, pp. 583–596, 2002.

[258] C. C. McAndrew and P. A. Layman, "MOSFET Effective Channel Length, Threshold Voltage, and Series Resistance Determination by Robust Optimization", IEEE Trans. Elec. Dev., Vol. 39, pp. 2298-2311, 1992.

[259] K. K. Ng, J. R. Brews, "Measuring the effective channel length of MOSFETs", IEEE Circ. And Dev. Mag., Vol. 6, no. 6, pp. 33-38, 1990.

[260] C. C. McAndrew, P. A. Layman, "MOSFET effective channel length, threshold voltage and series resistance determination by robust optimization", IEEE Trans. Elec. Dev., Vol. 39, no. 10, pp. 2298-2311, 1992.

[261] P. I. Suciu, R. L. Johnston, "Experimental Derivation of the Source and Drain Resistance of MOS Transistors", IEEE Trans. Elec. Dev. Vol. ED27, no. 9, pp. 1846-1848, 1980.

[262] B. Cabon-Till, G. Ghibaudo, S. Cristoloveanu, "Influence of source drain series resistance on MOSFET Field-effect mobility", Electronics Letters, Vol. 21, no. 11, pp. 457-458, 1985.

[263] C. Hao, B. Cabon-Till, S. Cristoloveanu and G. Ghibaudo, 'Experimental determination of short-channel MOSFETs parameters', Solid-State Electronics, Vol. 28, no. 10, pp. 1025-1030, 1985.

[264] F. Monsieur, Y. Denis, D. Rideau, J. Lacord, V. Quenette, G. Gouget, C. Tavernier, H. Jaouen, 'The importance of the spacer region to explain short channels mobility collapse in 28 nm Bulk and FD-SOI technologies', IEEE Proc. on ESSDERC 2014.

[265] G. J. Hu, C. Chang, Y. T. Chia, "Gate-Voltage-Dependent Effective Channel Length and Series Resistance of LDD MOSFET's", IEEE Trans. Elec. Dev. Vol. ED34, no. 12, pp. 2469-2475, 1987.

[266] K. Terada and H. Muta, "A new method to determine effective MOSFET channel length", Japan. J. Appl. Phys., Vol. 18, no. 5, pp. 953, 1979.

[267] J. G. J. Chem, P. Chang, R. F. Motta, and N. Godinho, "A new method to determine MOSFET channel length", IEEE Elec. Dev. Lett., Vol. EDL-l, no. 9, pp. 170, 1980.

[268] F. H. De La Moneda, H. N. Kotecha and M. Shatzkes, "Measurement of MOSFET Constants", IEEE Elec. Dev. Lett., Vol. EDL-3, no.1 , pp. 10-12, 1982.

[269] K. L. Peng and M. A. Afromowitz, "An Improved Method to Determine MOSFET Channel Length", IEEE Elec. Dev. Lett., Vol. EDL- 3, no. 12, pp. 360-362, 1982.

[270] B. J. Sheu, C. Hu, P. K. KO and F. C. Hsu, "Source-and-Drain Series Resistance of LDD MOSFET's", IEEE Elec. Dev. Lett., Vol. EDL-5, no. 49, pp. 365-367, 1984.

[271] K. –L. Peng, S. –Y. Oh. M. A. Afromowitz and J. L. Moll, "Basic Parameter Measurement and Channel Broadening Effect in the Submicrometer MOSFET", IEEE Elect. Dev. Lett., Vol. EDL-5, no. 11, pp. 473-475, 1984.

[272] J. Whitfield, "A Modification on 'An Improved Method to Determine MOSFET Channel Length", IEEE Elect. Dev. Lett., Vol. EDL-6, no. 3, pp. 109-110, 1985.

[273] L. Chang and J. Berg, "A Derivative Method to Determine a MOSFET's Effective Channel Length and Width Electrically", IEEE Elec. Dev. Lett., Vol. EDL-7, no. 4, pp. 229, 1986.

[274] F. Balestra, I. Hafez, G. Ghibaudo, "A new method for the extraction of MOSFET parameters at ambient and liquid helium temperatures", Journal de physique, Colloque C4, supplement au no. 9,

Tome 49, pp 817-820, 1988.

[275] F. Balestra, I. Hafez, G. Ghibaudo, “Modeling of electron mobility in silicon MOS inversion and accumulation layers at liquid helium temperature”, Electronics letters, Vol. 26, no. 19, pp 1633-1635, 1990.

[276] A. Cros, S. Harrison, R. Cerutti, P. Coronel, G. Ghibaudo, H. Brut, “New extraction method for gate bias dependent series ressitance in nanometric double gate transistors”, Proc. IEEE ICMTS, Vol. 18, 2005.

[277] D. Fleury, A. Cros, H. Brut, G. Ghibaudo, “New Y –Function-Based Methodology for Accurate Extraction of Electrical Parameters on Nano-Scaled MOSFETs”, Proc. IEEE ICMTS, 2008.

[278] D. Fleury, A. Cros, G. Bidal, J. Rosa, G. Ghibaudo, “A New Technique to Extract the Source/Drain Series Resistance of MOSFETs”, IEEE Elec. Dev. Lett., Vol. 30, no. 9, pp 975-977, 2009.

[279] N. Subramanian, G. Ghibaudo, M. Mouis, “Parameter Extraction of Nano-Scale MOSFETs Using Modified Y Function Method”, Proc. of IEEE ESSDERC, 2010.

[280] Y.Taur et al. “A New “Shift and Ratio” Method for MOSFET Channel-Length Extraction”, IEEE Elect. Dev. Lett., EDL-13(5), p. 267, 1992.

[281] S. Biesemans, M. Hendriks, S. Kubicek and K. De Meyer, “Accurate determination of Channel Length, Series Resistance and Junction Doping Profile for MOSFET 209ptimization in deep submicron technologies”, Proc. Symp. VLSI Technology Tech. Digest, p. 166, 1996.

[282] F. J. G. Sanchez, A. Ortiz-Conde, A. Cerdeira, M. Estrada, D. Flandre, J. J. Liou, “A method to extract mobility degradation and total series resistance of fully depleted SOI MOSFETs”, IEEE Trans. Elec. Dev., Vol 49, no. 1, 2002.

[283] K. O. Jeppson, “Static characterization and parameter extraction in MOS transistors”, Microelectronic engineering, Vol. 40, pp. 181-186, 1998.

[284] P. R. Karlsson and K. O. Jeppson, IEEE Trans. on Semiconductor Manufacturing, Vol. 9, pp. 215-222, 1996.

[285] H. Brut, A. Juge, and G. Ghibaudo, “New approach for the extraction of the gate voltage dependent series resistance and channel length reduction in CMOS transistors.”, Proc. of ICMTS, pp. 188-193, 1997.

[286] K. Yamaguchi, H. Asimiro, M. Yamawaki, and S. Asai, “A new variational method to determine effective channel length and series resistances of MOSFET’s”, in Proc. of ICMTS, pp. 123-126, 1997.

[287] Y. Denis, F. Monsieur, G. Ghibaudo, J. Mazurier, E. Josse, D. Rideau, C. Charbuillet, C. Tavernier, H. Jaouen, “New compact model for performance and du procédé variability assessment in 14nm FD-SOI CMOS technology”, IEEE Proc. of ICMTS, pp. 59-64, 2015

[288] C-L. Lou, W-K. Chim, D. S-H. Chan Y. Pan, “A novel single-device DC method for extraction of the effective mobility and source drain resistances of fresh and hot carrier degraded drain engineered MOSFET’s”, IEEE Trans. Elec. Dev., Vol. 45, no. 6, 1998.

[289] J. Kim, J. Lee, I. Song, Y. Yun, J. D. Lee, B-G. Park, H. Shin, “Accurate extraction of effective channel length and source/drain resistance on ultrashort channel MOSFETs by iteration method”, IEEE Trans. Elec. Dev., Vol. 55, no. 10, 2008.

[290] F. andrieux, “Transistors CMOS decananometriques à canaux contraints sur silicium massif ou sur SOI. Fabrication, caracterisation et etude du transport”, Ph.D. dissertation, INPG, EEATS, Grenoble, France

[291] PricewaterhouseCoopers ©, “Evolving landscape of technology deals: Semiconductor Industry-Device deal trends”, Technology Institute, 2015, available at https://www.pwc.com/us/en/technology/publications/assets/semiconductor-industry-device-deal-trends.pdf .

[292] D. Rosso, “Global Semiconductor Industry Posts Record Sales in 2014” , semiconductor industry association, February 2, 2015

[293] “Global semiconductor sales from 1988 to 2014 (in billion U.S. dollars)”, Statistica ©, available at http://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/.

[294] Rob van der Meulen, Janessa Rivera, “Gartner Says Worldwide Semiconductor Sales Expected to Reach $358 Billion in 2015, a 5.4 Percent Increase From 2014”, Gartner, Stamford, January 14, 2015

[295] KPMG, “KPMG global semiconductor survey – cautious optimism continues”, 2014, available at: https://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/semiconductor-survey-2014.pdf

[296] P. Goos, B. Jones, Optimal Design of Experiments: A Case Study Approach, Wiley, 2011

[297] Yuping Wu, “Parallel hybrid evolutionary algorithm based on chaos-GA-PSO for SPICE model parameter extraction”, Proc. on Intelligent Computing and Intelligent Systems, pp. 688 – 692, 2009

[298] Yiming Li, Yen-Yu Cho, “Parallel Genetic Algorithm for SPICE Model Parameter Extraction”, Proc. on Parallel and Distributed Du procédéing Symposium, 2006.

[299] W. Huyer, A. Neumaier , “Global Optimization by Multilevel Coordinate Search”, Journal of Global Optimization, Vol. 14, no. 4, pp 331-355, 1999.

[300] Panos M. Pardalos, H. Edwin Romeijn, Handbook of global optimization, USA, Florida, Kluwer Academic Publishers

[301] Gui-Bo Ye, Yifei Chen, Xiaohui Xie, “Efficient variable selection in support vector machines via the alternating direction method of multipliers”, Journal of Machine Learning Research, Vol. 15, pp. 832-840, 2011

[302] Ji Zhu, Hui Zou, “Variable Selection for the linear support vector machine”, Studies in Computational Intelligence, Vol. 35, pp. 35–59, 2007

[303] A. Rakotomamonjy, “Variable Selection Using SVM-based Criteria”, Journal of Machine Learning Research, Vol. 3, pp. 1357-1370, 2003

[304] Xiang Zhang, Yichao Wu, Lan Wang, Runze Li, “Variable selection for support vector machines in moderately high dimensions”, Journal of the Royal Statistical Society: Series B, Vol. 78, no. 1, pp 53–76, 2016.

[305] R. May, G. Dandy, H. Maier, “Review of Input Variable Selection Methods for Artificial Neural Networks”, in Artificial Neural Networks - Methodological Advances and Biomedical Applications, Ch. 2, pp. 19-44, 2011

[306] H. G. Mohammadi, P-E. Gaillardon, M. Yazdani, G. De Micheli, “A fast TCAD-based methodology for variation analysis of emerging nano-devices”, IEEE Internationnal Symposium on Defect and Faults Tolerance in VLSI and Nanotechnology Systems (DFTS), pp. 83-88, 2013.

[307] K. Kakehi, H. Aikawa, T. Tadokoro, H. Eguchi, T. Hirayu, H. Yoshimura, T. Asami, K. Ishimaru, “An efficient manufacturing technique based on du procédé compact model to reduce characteristic variation beyond du procédé limit for 40nm node mass production”, Symposium on VLSI Technology, pp.90-91, 2011.

[308] C. Gershenson, “Artificial Neural Networks for Beginners”, available at http://arxiv.org/ftp/cs/papers/0308/0308031.pdf

[309] R. Rojas, “Neural Networks: A Systematic Introduction. Springer, Berlin, 1996.

[310] C. Cortes, V. Vapnik, "Support-vector networks", Machine Learning, Vol. 20, no. 3, p. 273, 1995

[311] P. Goos, B. Jones, *Optimal Design of Experiments: A Case Study Approach*, Wiley, 2011

[312] Q. Zhou, W. Yao, W. Wu, X. Li, Z. Zhu and G. Gildenblat, "Parameter extraction for the PSP MOSFET model by the combination of genetic and Levenberg-Marquardt algorithms", In Proc. IEEE ICMTS, pp. 137-142, 2009

[313] R.A. Thakker, M.B. Patil, K.G. Anil, "Parameter extraction for PSP MOSFET model using hierarchical particle swarm optimization", Engineering Applications of Artificial Intelligence, Vol. 22, no. 2, pp. 317-328, 2009.

[314] Yuping Wu, "Parallel hybrid evolutionary algorithm based on chaos-GA-PSO for SPICE model parameter extraction", Proc. on Intelligent Computing and Intelligent Systems, pp. 688 – 692, 2009

[315] Yiming Li, Yen-Yu Cho, "Parallel Genetic Algorithm for SPICE Model Parameter Extraction", Proc. on Parallel and Distributed Du procédéing Symposium, 2006.

[316] K. Kakehi, H. Aikawa, T. Tadokoro, H. Eguchi, T. Hirayu, H. Yoshimura, T. Asami, K. Ishimaru, "An efficient manufacturing technique based on du procédé compact model to reduce characteristic variation beyond du procédé limit for 40nm node mass production", Symposium on VLSI Technology, pp.90-91, 2011.

[317] W. Huyer, A. Neumaier , "Global Optimization by Multilevel Coordinate Search", Journal of Global Optimization, Vol. 14, no. 4, pp 331-355, 1999.

[318] Panos M. Pardalos, H. Edwin Romeijn, Handbook of global optimization, USA, Florida, Kluwer Academic Publishers

[319] Gui-Bo Ye, Yifei Chen, Xiaohui Xie, "Efficient variable selection in support vector machines via the alternating direction method of multipliers", Journal of Machine Learning Research, Vol. 15, pp. 832-840, 2011

[320] Ji Zhu, Hui Zou, "Variable Selection for the linear support vector machine", Studies in Computational Intelligence, Vol. 35, pp. 35–59, 2007

[321] A. Rakotomamonjy, "Variable Selection Using SVM-based Criteria", Journal of Machine Learning Research, Vol. 3, pp. 1357-1370, 2003

[322] Xiang Zhang, Yichao Wu, Lan Wang, Runze Li, "Variable selection for support vector machines in moderately high dimensions", Journal of the Royal Statistical Society: Series B, Vol. 78, no. 1, pp 53–76, 2016.

[323] R. May, G. Dandy, H. Maier, "Review of Input Variable Selection Methods for Artificial Neural Networks", in *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, Ch. 2, pp. 19-44, 2011

[324] G. Rech, T. Terasvirta, R. Tschernig, "A simple variable selection technique for nonlinear model", Communications in Statistics - Theory and Methods, Vol. 30, no. 6, pp. 1227-1241, 2001.

[325] L. Rosas, M. Santoro, S. Mosci, A. Verri, S. Villa, "A Regularization Approach to Nonlinear Variable Selection", Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS), pp.653-660, 2010.

[326] Z. Lva, H. Zhua, K. Yub, "Robust variable selection for nonlinear models with diverging number of parameters", Statistics & probability letters, Vol. 91, pp. 90-97, 2014.

[327] S. Wu, H. Xue, Y. Wu, H. Wu, "Variable Selection for Sparse High-Dimensional Nonlinear Regression Models by Combining Nonnegative Garrote and Sure Independence Screening", Statistica Sinica, Vol.

24, pp. 1365-1387, 2014.

[328] P. Radchenko, G. M. James, “Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions”, Journal of the American Statistical Association, Vol. 105, no. 492, pp. 1541-1553, 2010.

[329] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician, Vol. 46, no. 3, pp. 175-185, 1992

[330] C. Gershenson, “Artificial Neural Networks for Beginners”, available at http://arxiv.org/ftp/cs/papers/0308/0308031.pdf

Abstract:

Recently, the race for miniaturization has seen its growth slow because of technological challenges it entails. These barriers include the increasing impact of the local variability and processes from the increasing complexity of the manufacturing process and miniaturization, in addition to the difficult of reducing the channel length. To address these challenges, new architectures, very different from the traditional one (bulk), have been proposed. However these new architectures require more effort to be industrialized. Increasing complexity and development time require larger financial investments. In fact there is a real need to improve the development and optimization of devices. This work gives some tips in order to achieve these goals. The idea to address the problem is to reduce the number of trials required to find the optimal manufacturing process. The optimal process is one that results in a device whose performance and dispersion reach the predefined aims. The idea developed in this thesis is to combine TCAD tool and compact models in order to build and calibrate what is called PCM (Process Compact Model). PCM is an analytical model that establishes linkages between process and electrical parameters of the MOSFET. It takes both the benefits of TCAD (since it connects directly to the process parameters electrical parameters) and compact (since the model is analytic and therefore faster to calculate). A sufficiently robust predictive and PCM can be used to optimize performance and overall variability of the transistor through an appropriate optimization algorithm. This approach is different from traditional development methods that rely heavily on scientific expertise and successive tests in order to improve the system. Indeed this approach provides a deterministic and robust mathematical framework to the problem. The concept was developed, tested and applied to transistors 28 and 14 nm FD-SOI and to TCAD simulations. The results are presented and recommendations to implement it at industrial scale are provided. Some perspectives and applications are likewise suggested.

Résumé:

Récemment, la course à la miniaturisation a vu sa progression ralentir à cause des défis technologiques qu'elle implique. Parmi ces obstacles, on trouve l'impact croissant de la variabilité local et process émanant de la complexité croissante du processus de fabrication et de la miniaturisation, en plus de la difficulté à réduire la longueur du canal. Afin de relever ces défis, de nouvelles architectures, très différentes de celle traditionnelle (bulk), ont été proposées. Cependant ces nouvelles architectures demandent plus d'efforts pour être industrialisées. L'augmentation de la complexité et du temps de développement requièrent de plus gros investissements financier. De fait il existe un besoin réel d'améliorer le développement et l'optimisation des dispositifs. Ce travail donne quelques pistes dans le but d'atteindre ces objectifs. L'idée, pour répondre au problème, est de réduire le nombre d'essai nécessaire pour trouver le processus de fabrication optimal. Le processus optimal est celui qui conduit à un dispositif dont les performances et leur dispersion atteignent les objectifs prédéfinis. L'idée développée dans cette thèse est de combiner l'outil TCAD et les modèles compacts dans le but de construire et calibrer ce que l'on appelle un PCM (Process Compact Model). Un PCM est un modèle analytique qui établit les liens entre les paramètres process et électriques du MOSFET. Il tire à la fois les bénéfices de la TCAD (puisqu'il relie directement les paramètres process aux paramètres électriques) et du modèle compact (puisque le modèle est analytique et donc rapide à calculer). Un PCM suffisamment prédictif et robuste peut être utilisé pour optimiser les performances et la variabilité globale du transistor grâce à un algorithme d'optimisation approprié. Cette approche est différente des méthodes de développement classiques qui font largement appel à l'expertise scientifique et à des essais successifs dans le but d'améliorer le dispositif. En effet cette approche apporte un cadre mathématique déterministe et robuste au problème. Le concept a été développé, testé et appliqué aux transistors 28 et 14 nm FD-SOI ainsi qu'aux simulations TCAD. Les résultats sont exposés ainsi que les recommandations nécessaires pour implémenter la technique à échelle industrielle. Certaines perspectives et applications sont de même suggérées.